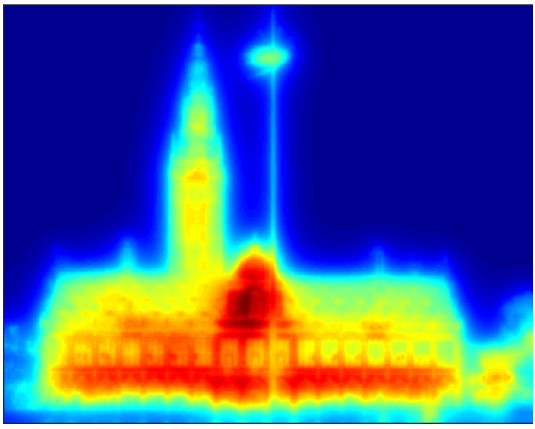
KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Convolution Based Modeling Methodology for the Fast Thermal Analysis of 3D-ICs



Federica Lidia Teresa Maggioni

Dissertation presented in partial fulfillment
of the requirements for the degree of
Doctor of Engineering Science (PhD):
Mechanical Engineering

Supervisor:
Prof. dr. ir. M. Baelmans
Prof. dr. I. De Wolf, co-supervisor

May 2016

Convolution Based Modeling Methodology for the Fast Thermal Analysis of 3D-ICs

Federica Lidia Teresa MAGGIONI

Examination committee:
Prof. dr. ir. W. Sansen, chair
Prof. dr. ir. M. Baelmans, supervisor
Prof. dr. I. De Wolf, co-supervisor
Prof. dr. ir. D. Vandepitte
Dr. ir. H. Oprins (IMEC)
Prof. dr. ir. J. Vandewalle
Dr. ir. E. Beyne (IMEC)
Prof. dr. ir. M. Rencz (Technical University of Budapest)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Mechanical Engineering

In collaboration with

imec vzw

Interuniversitair Micro-Elektronica Centrum vzw
Kapeldreef 75, 3001 Leuven, Belgium

May 2016

Preface

Once upon a time.... all good stories start like this, and the story of this PhD started, almost by chance, four and a half years ago when my father came back home from one of his business trips: "The son of a colleague of mine is doing a PhD at imec and he's really happy with it, why don't you look at their website?" That was the first time I ever heard about imec and Leuven. So, thanks papà and Alex for having offered me the very early starting point of this experience. Also Sara should be mentioned as one of the first people having a role in the preface of my Leuven adventure: she was already living here, she enthusiastically described the city and she offered me a place to sleep when I came to look for an accommodation.

The next people who appear in this story are dr. Herman Oprins and prof. Ingrid De Wolf who, after a phone call, decided to offer me the possibility to start working on thermal modeling. Shortly after my arrival in Leuven, also prof. Tine Baelmans became a very relevant part of this PhD. I really need to thank these three people for their help, guidance, support and suggestions every time I needed a hint on how to proceed. They have always been ready to help me, with a nice smile on their faces. A special thanks to Herman, whose daily advices have been really useful to be able to achieve this final goal.

I also want to thank my assessors and the jury members for their effort in reading the text and in helping me to improve the quality of my manuscript with their valuable comments during the private defense. I did my best to take all of them into account. Also from a technical point of view a big thanks to dr. Vladimir Cherman, who provided me with all the experimental results needed to validate the model, and to all my colleagues in REMO and TME, with whom I had the pleasure to have both technical and non-technical discussions.

Four years seemed a very long time at the beginning and it has not always been easy, neither from a technical nor from a daily life perspective: the decision to leave family, a boyfriend, friends and a fixed (but quite boring) job for something completely unknown 1000 km away has not been an easy one. The adaptation to the new environment took also quite some time before I could feel "at home" in Leuven. A really big thanks goes to all the people who convinced me to start this adventure and who continued supporting me from the beginning to the end,

especially in the gloomy days: papà, mamma, Ale, Paolo, Giaime,...

I also want to thank all the people who helped me to familiarize with this new country and new city: Caroline, her family and my (more-or-less) distant Belgian relatives. A big thanks also to all the SLOK people with whom I enjoyed really nice runnings on Tuesday evenings and orienteering races on Sundays. A special thanks for this goes to everyone who drove me all around the country, to the most unknown (but really nice) places where orienteering races normally take place: An, Miek, Hans, Bruno, Jiri, Vendy, Dries, Ariane, Sara, Sofie... Having found all these nice people in Leuven allowed me to continue with my favorite hobby/sport and to feel "more at home". Orienteering friends in Italy also deserve a special thanks, I mention Maria Chiara just to name one: every time I went back to Italy, they always welcomed me as if I never even left.

My friends and colleagues with whom I had a nice time during lunch breaks, evenings and weekends deserve also a special mention: Nabi, Vice, Andrej, Anna, Sofie, Sanja, Luka, Kristof. Together with the other people in REMO, they provided a pleasant and friendly environment to work in and to relax during breaks and free time.

Sara, Marco and An deserve a special paragraph in this thank you page just for them. Dinners, board games, movies, walks, runnings, bike trips, travels, chatting, Christmas markets,...: you made my time in Leuven much better and you really made me feel "at home"!

Another special paragraph is for Giaime, who didn't stop me from having this experience and who came over to learn Dutch and to try to get a job over here. I also need to thank him for his continuous support and patience in the bad days.

All in all, even with its ups and down, this has been a great experience from which I learned a lot, or even more than that. On top of the technical expertise, which I gained during the "working hours" and over which you can read more in the rest of this book, the knowledge of a new language and of a new culture(s) made me a different person than who I was four years ago. I really feel like I changed and I learned a lot in this period, also thanks to the multicultural environment in Leuven and to the really nice and helpful people I met here.

Once again, I need to thank all the main actors of this story, which I've already mentioned above, but also all the other people with whom I enjoyed nice moments or chats and that I'll preserve in my memories. As in all stories, there have been first actors and background actors, but the story of this PhD would have not been the same even without only one of them.

Now it is time for a new (still unknown) challenge but I'll keep with me all the valuable teachings and good memories of my time in Leuven.

Federica
Leuven, 12th April 2016

Abstract

The relevance of accurate predictions of the thermal behavior of microelectronic systems has been increasing since the introduction of 3D-ICs. Due to the vertical stacking of the active dies the reliability issues related to high temperature and temperature gradients are, indeed, exacerbated. Different thermal modeling strategies have been developed with the aim of providing quick estimations of the device temperature under operating conditions. It is, indeed, important to be able to quickly compare the thermal impact of different design and technological parameters already during the design phase.

In this thesis, an *easy-to-use* fast thermal model methodology based on the Green's function theory is presented. It provides *highly resolved* temperature maps on selected levels (or selected points), avoiding the calculation in locations that are not thermally significant and reducing, therefore, the computational time. The model is able to deal with both the steady state and the transient regime and it proved to be two orders of magnitude *faster* than conventional finite element methods, maintaining the error on peak temperatures below *5%*.

The core of the algorithm is constituted by the convolution between 1) the thermal response of the system to localized and impulsive power dissipation and 2) the actual dissipated power map. However, this basic convolution approach is valid only for stack (layered) structures in which multiple layers, of homogeneous material and with infinite horizontal size, are placed on top of each other. The "method of images" is used to take into account the *finite dimension* of the stack while correction strategies are applied to account for the thermal impact of specific *μbump layouts* (only in case of a two dies stack in the steady state regime) and of the *package*. Moreover, an a posteriori mathematical transformation, the Kirchhoff transformation, has been proposed to deal with the *temperature dependency* of the silicon thermal conductivity. By overcoming the limitations of the basic convolution approach, the developed fast thermal model is able to deal with realistic 3D-IC configurations. This has been proved by a successful experimental validation with respect to measurement data.

The model has also been extended to deal with other geometries commonly available in microelectronic applications (side by side integration on an interposer

and stack of dies with different sizes). Moreover, to demonstrate the *applicability* and the *easiness-of-use* of the developed methodology, the model has been applied to perform realistic analyses that might be needed during the design phase of an IC.

It is, therefore, concluded that the developed fast thermal model can be a valid alternative to conventional thermal modeling strategies for 3D-ICs and related geometries: the computational time is, indeed, strongly reduced and high accuracy is maintained.

Beknopte samenvatting

Sinds de introductie van de driedimensionaal gestapelde chipverpakkingen, is het belang van de nauwkeurige voorspelling van het thermische gedrag van elektronische componenten sterk toegenomen. De verticale stapeling van actieve componenten leidt immers tot hogere temperaturen en/of hogere temperatuurgradiënten en bijgevolg mogelijk tot meer betrouwbaarheidsproblemen. Daardoor is het belangrijk om op een snelle manier de thermische impact van verschillende ontwerpvariabelen te kunnen vergelijken. Verschillende thermische modelleringsstrategieën zijn ontwikkeld om een snelle en nauwkeurige voorspelling te maken van de werkingstemperatuur van de component onder bedrijfsomstandigheden.

In deze doctoraatsthesis wordt een *gebruiksvriendelijke* en snelle thermische modeleringsmethode voorgesteld, die gebaseerd is op de Greense functie. De voorgestelde methode is in staat de temperatuurverdeling op geselecteerde niveaus (of geselecteerde punten) te voorspellen. Omdat hierbij de berekening voor niet-relevante locaties vermeden wordt, kan de berekeningstijd significant gereduceerd worden. De voorgestelde methode is in staat om zowel de regimetoestand als het transiënte thermische gedrag voorspellen. In de thesis wordt aangetoond dat de voorgestelde berekeningsmethode tot twee grootteordes *sneller* is dan conventionele eindige-elementenmethoden en dat de fout op de piektemperatuur kleiner dan 5% is.

De kern van het algoritme bestaat uit de convolutie tussen de thermische respons van het systeem op een impulsieve vermogensdissipatie enerzijds en de werkelijke vermogensverdeling over het oppervlak van de actieve chip anderzijds. Deze op convolutie gebaseerde basismethode is echter alleen geldig voor gestapelde (gelaagde) structuren waarin meerdere lagen, van homogeen materiaal en met oneindige horizontale afmetingen, boven op elkaar geplaatst worden. De "methode van de gespiegelde bronnen" wordt gebruikt om rekening te houden met de *eindige afmetingen* van de stapeling terwijl correctiestrategieën worden toegepast om de thermische invloed van *specifieke patronen van microverbindingen* (alleen maar voor stapelingen van twee chips in de regimetoestand) en van de *chipverpakking* in rekening te brengen. Bovendien wordt een a posteriori wiskundige transformatie,

de Kirchhoff transformatie, voorgesteld om de *temperatuurafhankelijkheid* van de thermische geleidbaarheid in rekening te brengen. Door de beperkingen van de op convolutie gebaseerde basismethode te overwinnen, is het ontwikkelde snelle thermische model in staat het thermische gedrag van realistische 3D-ICs nauwkeurig te analyseren. Het model is bovendien succesvol experimenteel gevalideerd aan de hand van meetdata.

Het model is verder uitgebreid om ook andere geometrieën die veelvuldig in de micro-elektronica voorkomen te kunnen behandelen zoals naast elkaar geplaatste chips op een interposer en een stapeling van chips met verschillende afmetingen. Om de *toepasbaarheid* en de *gebruiksvriendelijkheid* van de ontwikkelde methode te demonstreren, is het model vervolgens toegepast voor realistische analyses die tijdens de ontwerpfase van een chip kunnen voorkomen.

Daardoor kan er besloten worden dat de ontwikkelde snelle thermische modeleringsmethode een bruikbaar alternatief kan zijn voor conventionele thermische modeleringsstrategieën voor 3D chipverpakkingen en verwante geometrieën: de berekeningstijd is immers significant verminderd terwijl de hoge nauwkeurigheid bewaard blijft.

Abbreviations

BC	Boundary condition
BEOL	Back end of line
CTM	Compact thermal model
DFT	Discrete Fourier transform
DOE	Design of experiments
F2B	Face-to-back
F2F	Face-to-face
FD	Finite difference
FEM	Finite element method
FFT	Fast Fourier transform
FTM	Fast thermal model
GCI	Grid convergence index
HP	High power
HS	Hot spot
HSR	Hot spot response
IC	Integrated circuit
IDFT	Inverse discrete Fourier transform
LP	Low power
PCB	Printed circuit board
PDE	Partial differential equation
PM	Power map
POD	Proper orthogonal decomposition
RC	Resistance-capacitance

TSV Through silicon via

List of Symbols

(x_{HS}, y_{HS})	Position of the HS (m, m)
$*_{2D}$	Convolution operator in 2D
$*_{3D}$	Convolution operator in 3D
$/^{(*)}$	Inverse 2D-convolution operator
α	Matrix of weights, based on μbumps location, used in the inclusion of the μbumps thermal impact in the FTM
$*$	Convolution operator
$\bar{C}$	Correction profile for including the package thermal effect in the steady state regime
$\bar{C}_l$	Correction profile for including the package thermal effect in transient regime: effect l time steps after impulsive power dissipation
$\bar{C}_{int,ij}$	Correction profiles, in case of interposer geometry, on die j due to power dissipation on the active die i
$\bar{T}$	Expected average temperature value during chip operation (°C)
$\bar{t}_{ss}$	Number of time steps needed by the HSRs to reach steady state, according to the time step used in the FTM
$\bar{\Theta}_{z_i}(i, j, z_j, t_k; t_l)$	Discrete temperature increase at time t_k and on level z_j due to impulsive power dissipated at time $t_k - t_l$ on level z_i (°C)
$\bar{\theta}_{z_i}(x, y, z_j, t; t_0)$	Temperature increase at time t and on level z_i due to impulsive power dissipated at time $t - t_0$ on level z_i (°C)
$\bar{h}$	Grid size, edge of the grid cell (m)
$\bar{h}_{HS}$	Size of the HS, edge of the grid cell (m)

$\bar{t}_f$	Number of simulated time steps
$\boldsymbol{C}$	Capacitance matrix
$\boldsymbol{G}$	Conductance matrix
$\boldsymbol{p}$	Vector storing the known informations in a linear system
$\boldsymbol{q}$	Vector storing the local heat flux density (W/m^2)
$\boldsymbol{T}$	Temperature vector
$\boldsymbol{x}$	Three dimensional space variable (m, m, m)
$\boldsymbol{\xi}$	Generic variable
Δt	Time step used in the transient FTM (s)
γ	Matrix of weights, based on fitting, used in the inclusion of the μbumps thermal impact in the FTM
$\hat{T}$	Transformed temperature field via Kirchhoff transformation (°C)
κ	Thermal diffusivity, $\kappa = k/\rho c$ (m^2/s)
$\lceil x \rceil$	Smallest integer number greater than or equal to x
$\lfloor x \rfloor$	Greatest integer number smaller than or equal to x
$\mathcal{F}(\boldsymbol{x})$	Fourier transform of $\boldsymbol{x}$
$\mu bEff$	Maximum temperature reduction on the top die, with respect to the case of uniform underfill material, when a non-uniform μbump array layout is used in the interface layer. Uniform power dissipation on top die (°C)
μBumpsMap	Binary matrix representing the μbump layout: 1 indicates a μbump cell and 0 an underfill cell
Ω	Spatial domain in which the PDE is defined
$\partial\Omega$	Boundary of the spatial domain in which the PDE is defined
ρ	Mass density (kg/m^3)
σ	Sensitivity of the diode ($V/°C$)
τ	Time constant of the system (s)
θ	Temperature increase (°C)
$\Theta_{FEM,unif,ij}$	Temperature increase, computed by FEM, on die j due to uniform power dissipation on die i in the interposer geometry (°C)

$\Theta_{FTM,pack}$	Temperature increase profile obtained by the package FTM (°C)
$\Theta_{FTM,stack}$	Temperature increase profile obtained by the stack FTM (°C)
$\Theta_{FTM1,unif,ij}$	FTM for interposer, temperature increase computed by the stack FTM on die j due to uniform power dissipation on die i in the stack configuration considered in *FTM1* (°C)
$\Theta_{FTM2,unif,ij}$	FTM for interposer, temperature increase computed by the stack FTM on die j due to uniform power dissipation on die i in the stack configuration considered in *FTM2* (°C)
$\Theta_{i,j}$	Discrete temperature increase on level z_j due to power dissipated on level z_i (°C)
$\Theta_{pack,unif}$	Steady state temperature increase profile obtained by FEM for uniform power dissipation in a package configuration (°C)
$\Theta_{stack,unif}$	Steady state temperature increase profile obtained for uniform power dissipation in a stack configuration (°C)
$\theta_{z_i}(x, y, z_j, t)$	Temperature increase on level z_j due to power dissipated on level z_i (°C)
$\tilde{C}$	Equivalent capacitance (W/mK)
$\tilde{h}(x, y, t)$	Equivalent heat transfer coefficient (W/m^2K)
$\tilde{k}_z$	Equivalent out-of-plane thermal conductivity (W/mK)
$\tilde{k}_{x,y}$	Equivalent in-plane thermal conductivity (W/mK)
$\tilde{u}(\boldsymbol{\xi})$	System response to unit perturbation
$\tilde{\alpha}$	Allowed error percentage
$\tilde{\rho}$	Ratio of the area covered by μbump arrays versus the total die area
A	Base area (m^2)
C	Thermal capacitance (J/K)
c	Specific heat capacity (J/kgK)
cs	Chip size (m)
Eff_1	Temperature reduction on the top die when μbump array equivalent material properties are used instead of the underfill ones. Uniform die-die interface material and uniform power dissipation on top die (°C)

Eff_2	Temperature reduction on the bottom die when underfill material properties are used instead of the μbump array equivalent ones. Uniform die-die interface material and uniform power dissipation on top die (°C)
$f(\boldsymbol{\xi})$	Known quantity in a linear PDE
F_d	Mask used to compute the α-weights matrix while including the μbumps thermal impact in the FTM
$G(\boldsymbol{x}, t; \boldsymbol{x}_0, t_0)$	Green's function at $(\boldsymbol{x}, t)$ due to impulsive perturbation at $(\boldsymbol{x}_0, t_0)$
$G_{z_i}(x, y, z_j, t)$	Green's function obtained on level z_j due to power dissipated on level z_i at $(x_0, y_0, t_0) = (0, 0, 0)$
H	Heaviside function
h	Heat transfer coefficient (W/m^2K)
h_b	Heat transfer coefficient on the bottom boundary (W/m^2K)
h_t	Heat transfer coefficient on the top boundary (W/m^2K)
$h_{b,pack}$	Heat transfer coefficient on the bottom surface of the package configuration (W/m^2K)
$h_{t,pack}$	Heat transfer coefficient on the top surface of the package configuration (W/m^2K)
$HSR_{z_i}(d, z_j, t)$	HSR function on level z_j at time t and at a distance d from the HS center due to power dissipated on level z_i, ($°C/W$ in steady state, $°C/J$ in transient)
$impr$	Estimation of the maximum relative improvement achievable by applying the package correction
k	Material thermal conductivity (W/mK)
$k_0 = k(T_0)$	Thermal conductivity value at temperature $T = T_0$ (W/mK)
$k_{Si,Q}$	Silicon thermal conductivity depending on the total dissipated power (W/mK)
k_{Si}	Silicon thermal conductivity (W/mK)
L	Partial linear differential operator
l	Thickness (m)
N	Number of elements in the PM
N^e	Number of elements in the extended PM, PM^e

N_c	Number of columns in the PM
N_p	Number of layers in which power is dissipated
N_r	Number of rows in the PM
N_t	Number of layers in which the temperature profiles are computed
NI	Number of images per side
p	Order of grid convergence
$PM^e_{z_i}(x, y, t)$	Power map, extended with NI images per side, dissipated on level z_j at time t (W in steady state, J in transient)
$PM_{z_i}(x, y, t)$	Power map dissipated on level z_j at time t (W in steady state, J in transient)
Q	Total dissipated power (W)
q	Dissipated power density (W/m^2)
$Q_{int,unif,i}$	Total power dissipated on die i while computing $\Theta_{int,unif,ij}$ (W)
r	Grid refinement ratio
R_1	Conductive thermal resistance of die 1 (K/W)
R_2	Conductive thermal resistance of die 2 (K/W)
R_i	Conductive thermal resistance of the interface layer (K/W)
$R_{b,c}$	Convective thermal resistance on bottom of the die stack (K/W)
r_{HS}	Radius of the HS generating the HSRs (m)
$R_{t,c}$	Convective thermal resistance on top of the die stack (K/W)
R_{th}	Thermal resistance (K/W)
T	Temperature ($°C$)
t	Time variable (s)
t_f	Total simulated time (s)
T_{amb}	Ambient temperature ($°C$)
T_i	Temperature in location i ($°C$)
$T_{k_{Si}(T)}$	Temperature profile computed considering the temperature dependency of the silicon thermal conductivity ($°C$)

$T_{k_{Si}=120}$	Temperature profile computed assuming a fixed value of the silicon thermal conductivity equal to 120 W/mK ($°C$)
T_{ref}	Reference temperature ($°C$)
T_{ss}	Temperature at steady state ($°C$)
t_{ss}	Time needed by the HSRs to reach steady state (s)
$u(\boldsymbol{\xi})$	Solution of a linear PD
z_b	z-coordinate of the bottom boundary
Z_{th}	Impedance curve ($°C$)
z_t	z-coordinate of the top boundary

Contents

List of Figures

List of Tables

Part I

Model for the Stack Configuration

Chapter 1

Introduction

1.1 Introduction

The operating conditions of an integrated circuit (IC) are associated to power dissipation. However, not only the desired fulfillment of tasks, but also an unwanted and unavoidable temperature increase in the IC, is associated to this power dissipation. The thermal issues related to this phenomenon are presented in this Chapter. Different simplified modeling approaches have been developed to quickly forecast the temperature increase in the chips and prevent critical situations. An overview of the state of the art concerning fast thermal modeling (FTM) methodologies is also reported hereafter. At the end of the Chapter, the goals and the contributions of this thesis, with respect to existing approaches, are formulated.

1.2 Thermal analysis for 3D chip integration

1.2.1 Thermal issues in microelectronics

High temperatures and temperature gradients have a negative effect on ICs performance and reliability. A simple first order model, the Arrhenius model, shows, indeed, for the temperature accelerated failure modes, an exponential dependency of the mean time to failure on temperature [40]. In fact, the impact of a small variation in the operating temperature of just 10°C-15°C may be so drastic that the life time of the device is halved [26]. This strong relation is mainly due to several failure mechanisms that are accelerated and exacerbated at high

temperature and high temperature gradients. Examples of phenomena affected by temperature are listed below.

- *Electromigration* is a failure mechanisms associated with the gradual movement of material in electrical conductors and, as a consequence, can cause loss of connection and permanent failures of the devices. This phenomenon, which depends on temperature, is accelerated not only by high temperature values but also by high spatial temperature variations [51].
- The *leakage power* is another failure mechanism, related to an unwanted loss of energy from a charged capacitor, that causes an increase of power consumption. As leakage is exponentially related to temperature with a positive feedback, it may lead to extreme temperatures that may irreversibly damage the circuit [38].
- *Carrier mobility* also degrades at high temperature. This phenomenon negatively affects performance because the operating speed reduces.
- *RC delay* is, as well as carrier mobility, an issue related to the electro-thermal coupling [53]. It originates from an electrical resistance increase caused by a temperature increment and it is of particular concern in the interconnections. The problem is that a longer interconnect RC delay degrades performances and can cause logic failures. The difference in the resistivity of copper, for example, with increasing temperature from 20°C to 120°C is 39% [40]. In other words, every 20°C temperature increase causes a 5%-6% increment in RC delay in interconnections, meaning that clock skew problems become significant with 15°C-20°C temperature difference [26].
- *Package fatigue* and *plastic deformation* may also cause permanent failures. These phenomena originate from thermal cycling, which may occur from system power on/off cycles as well as from workload rate changes. Not only the magnitude of the dissipated power but also the cycling frequency affects the failure rate: for a 10°C increase in the magnitude of cycles a 16 times smaller mean time to failure can be expected [26].

Because both high temperatures and high temperature gradients are related to failures mechanisms, situations of particular concern are the ones in which these two conditions are combined. This occurs in case of localized high temperature peaks, namely *hot spots* (HSs).

The research towards smaller, faster and more powerful devices should deal with, and take into account, all these temperature related phenomena to be able to create reliable devices with high performances.

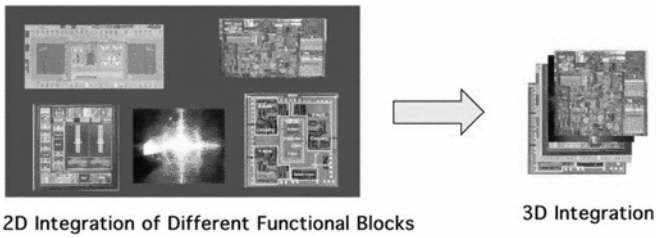


Figure 1.1: Integrated microsystem in 2D and 3D [14].

1.2.2 3D system integration technology

Higher functional density and higher performances are the main features driving the evolution and the research in the microelectronic industry. Up to now, industry could keep up with the market's demand by scaling, miniaturizing and using advanced IC packaging and integration techniques [14]. However, these approaches reach their physical limits in 2D technology and the further reduction of the device size becomes more and more challenging from both a technological and a financial perspective [41]. This led to the development of 3D-ICs in which active dies are vertically integrated.

The concept of 2D and 3D technology is illustrated in Figure 1.1. While in the 2D approach the different functional blocks are placed on the same level, meaning that they can be arranged and connected only along the horizontal direction, in the 3D approach they are divided over more dies that are stacked on top of each other. In this way, the vertical direction is exploited for integration. 3D technology has been considered as the key to keep up with *Moore's law*. According to this empirical law, presented in 1965, the number of transistors on an IC doubles approximately every two years [68] (Figure 1.2). Even if the formulation of Moore's law was based on observations over the history, it revealed to be accurate also for future times. This is mainly because it has been considered as a target in the microelectronic industry.

The use of the third dimension allows meeting most of the goals in IC development: miniaturization, integration of different technologies, small form factor and increased performances. Stacking the dies on top of each other allows, indeed, a large reduction of the form factor and, therefore, of the overall size of the final system. This is because the in-plane dimensions are significantly shortened with just a small increment in the out-of-plane one. Moreover, the density of integration is higher in the 3D approach because the different dies are vertically interconnected and not through the PCB. In this way, the interconnects length is shortened and, as a consequence, time delay is reduced and performances are increased. On top of this, it is also important to note that the parasitic losses in interconnections reduce by shortening them. Consequently, power consumption is also reduced [14].

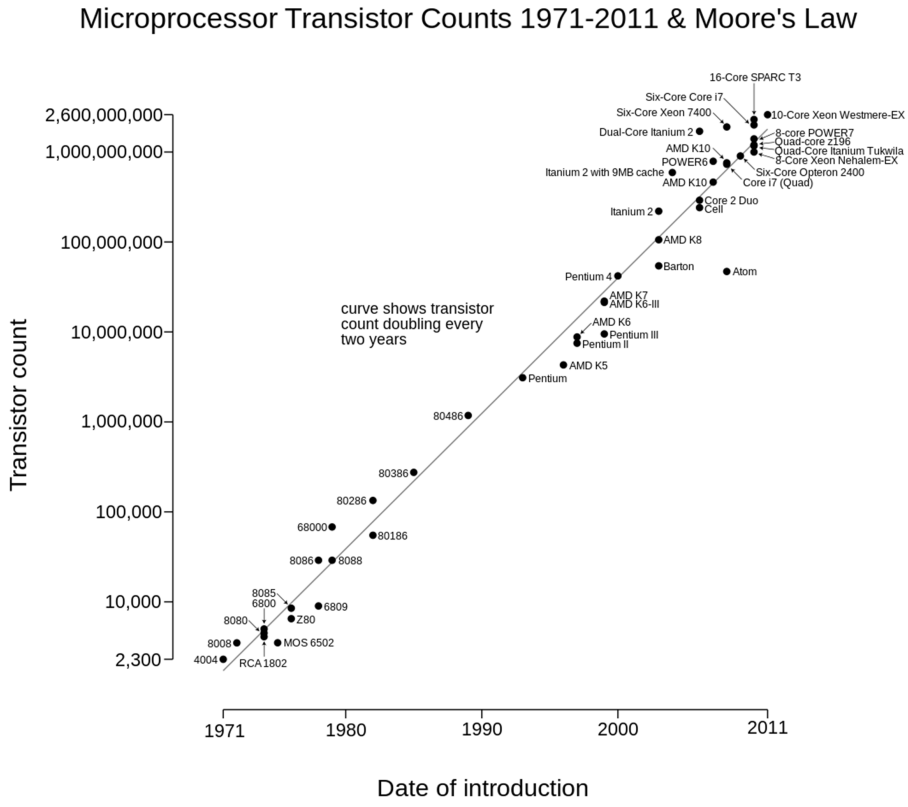


Figure 1.2: Transistor counts for integrated circuits plotted against their dates of introduction. The curve shows Moore's law - the doubling of transistor counts every two years [111].

Three main categories can be defined in 3D technology: *3D stacking of packages*, *3D chip stacking* and *3D on chip integration* [107].

- The *3D stacking of packages* (also *Package-on-Package* or *PoP*) is the direct extension of the single chip package strategy into 3D: it consists in the stacking of individual 2D packages. Two of the main advantages of this approach are that the number of elements to be assembled on the PCB is reduced and that, since only known-good-dies are used, the yield is high and the reliability issues low. If, on the one hand, this approach presents less technological issues than the next ones and it has already been implemented in different commercialized applications as mobile phones, cameras, MP3,. . . , it won't be able to fulfill the performance and miniaturization requirements forecast for the future.
- In *3D chip stacking* (also *stacked IC* or *3D-SIC*) the wafers are processed

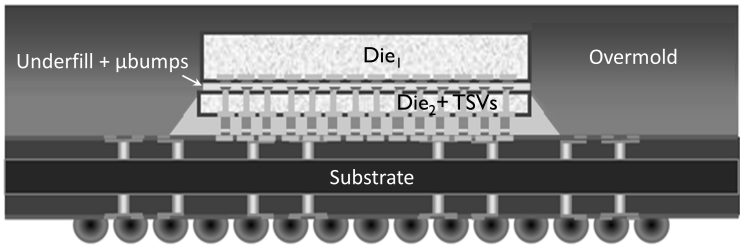


Figure 1.3: Schematic of a 3D die stack (from [75]).

separately and then bonded. The stacking can be done at a wafer-to-wafer, chip-to-wafer or chip-to-chip level. Through silicon vias (TSVs), which are copper vias etched in the silicon, are used to allow the electrical connections between the PCB and the upper layers, through the halfway silicon dies. Moreover, the stacked dies need to be bonded to each other. One option to achieve this aim is the use of metallic interconnects (μbumps). These structures have an approximately cylindrical shape and they are made of Cu, Sn and intermetallic compounds (Cu_3Sn and Cu_6Sn_5), which form during the thermo-compression bonding phase [73]. For mechanical reasons, the interconnects are surrounded by underfill material, which typically has low thermal conductivity [14] (cf. Figure 1.3). It is worth to note that the μbumps, not only provide mechanical support and electrical connections, but they are also thermally conductive, enhancing, in this way, the overall thermal conductivity. An important step in this technology, which has consequences from a thermal point of view, is the wafer thinning. The dies are, indeed, thinned down to allow for TSVs that are shorter and with a smaller diameter. In this way, process costs are reduced but, from a thermal point of view, the consequence is that the lateral thermal spreading in the dies is also reduced and the hot spots are enhanced.

- *3D on chip integration* (also *monolitic 3D*) is a truly homogeneous kind of integration where active device layers are built up subsequently on top of an initial layer. Many technological issues are coupled with this approach and, as a consequence, it is still in the R&D stage.

The technology considered in this work is the 3D chip stacking using TSVs and μbumps.

On top of the technological issues associated with the development and commercialization of the 3D-IC stacks, thermal management plays an important role. As already explained in Section 1.2.1, temperature related issues threaten the performances and reliability of the devices and they become even more pronounced in 3D technology. This is mainly because:

- more power is dissipated over the same area available for cooling;

- the dies are thinned down reducing the lateral thermal spreading;
- underfill material with low thermal conductivity is used to stack the dies.

For these reasons, this 3D-IC stacking technology is currently used in low power applications such as memory modules, logic-memory stacks, image sensors, etc. and research is being carried out towards more efficient thermal management and more advanced cooling solutions such as liquid cooling, jet impingement, inter-layers liquid cooling, intra-layer liquid cooling etc. [7, 12, 114].

1.2.3 Thermal modeling

In this frame, thermal modeling has a great importance in avoiding designs with too high temperature peaks and/or temperature gradients. This means that, besides mechanical and electrical constraints, also the thermal ones have to be taken into account in order to guarantee the required performances. The sooner all these issues are tackled, the fewer corrections are needed afterwards. From a thermal point of view, the terminology *thermal aware design* is used when the thermal constraints are considered during the design phase. For each specific design, different options for the geometry and the cooling strategy should be analyzed and compared in order to determine solutions that are acceptable from a thermal point of view. Moreover, a good layout and architecture of the functional blocks is essential to reduce thermal issues by placing, for example, high power dissipating modules far away form each other. Also a proper design of TSVs and μbumps layouts can help in improving the overall cooling: more μbumps increase, indeed, the thermal conductivity of the interface layers but, at the same time, the process costs (Chapter 4 and Section 9.4.2).

Different techniques are used for the thermal modeling of ICs, depending on the scope of the simulation. A first option is a full numerical approach, in which both the conduction (within the package) and the convection (in the fluid surrounding the ICs and/or in the microchannels used to cool down the device in case of advanced, high power applications) are explicitly modeled. This is most often performed by means of computational fluid dynamics (**CFD**), for which several commercial software are available. However, since the main interest in this thesis is directed towards the conduction within the package, the CFD approach won't be considered.

A second group, in which the interest is limited to conduction while the effect of convection is included just through appropriate Robin's boundary conditions, is typically treated by means of the finite element method (**FEM**). This is also a well established technique, for which several commercial software packages are available, and it is commonly used for scenarios similar to the ones considered in this thesis. Thermal FEM simulations are quite easy and they run faster than, for example, FEM models for structural analysis. This is because they have to

deal with just one degree of freedom (temperature) while, for structural analysis, the degrees of freedom are six (three rotations and three displacements). For fine enough meshes, the computed temperature profiles are very accurate but specific modeling expertise is required to create a good mesh that allows to obtain accurate results without using unneeded computational time [37]. However, fully detailed, 3D, transient models may take hours (or even days) to run [78]. In particular situations, typical for the design phase of microelectronic devices, this may become a relevant drawback of this approach. If, for example, the thermal impact of different geometrical/material parameters has to be compared, or if the positioning of multiple active blocks needs to be thermally optimized, or if electro-thermal simulations are required, the solution of the thermal models is expected to be obtained much faster, in the order of seconds or minutes. Moreover, if the thermal modeling is outsourced or if the people in charge of the layout design cannot share the proprietary information of the considered structure under development to the thermal engineers, the thermal FEM model becomes impossible to be run. This is because, in FEM, the whole structure has to be discretized and all the material properties are required.

Considering the limitations of the FEM approach in some situations, different research groups started looking into alternative options that allow obtaining relevant thermal related information more quickly (in the order of seconds or minutes) and more easily. These compact thermal models (**CTMs**) or computationally fast thermal models (**FTMs**) represent the third approach applicable to the thermal modeling of ICs. They significantly reduce the computational time but, to achieve this goal, further simplifications are needed, leading to a reduction in accuracy.

The first developed fast thermal modeling methodologies were limited to study the temperature profiles at *steady state*, i.e. after the system reached thermal equilibrium. If this kind of simulations is, on the one hand, much simpler and faster than the *transient* one, in which the time evolution of the temperature profile is modeled, it represents, on the other hand, a worst case scenario. Actual devices work, indeed, in dynamic thermal regime. Cores are subsequently switched on and off depending on the workload and the load can be moved from one core, when it reaches a problematic temperature, to another one in a colder location. This helps in maintaining the temperature within safe limits, which has a positive impact on speed and power consumption. Considering the results obtained from steady state simulations may lead to opt for a more advanced and expensive cooling solution than the one actually needed in real working conditions [19].

1.3 State of the art, fast thermal models

Transient and steady state fast thermal modeling methodologies for conduction heat transfer in electronic packages have been widely studied in recent years.

Various strategies have been presented in the literature, each of them focusing on solving particular modeling criticalities depending on the specific aim of the methodology itself. This variety of scopes results in some differences in the abilities and characteristics of the developed FTM. The main properties that characterize a FTM are:

- Physical or behavioral base;
- Need of finite element or measurement results upfront, before the FTM of a specific structure (fixed dimensions, package, materials,...) can be built;
- Capability to include multiple layers and heterogeneous materials;
- Level of details and granularity of the model (e.g. die stack, die stack and package, die stack with μbumps and/or TSVs, . . .);
- Possibility to model different configurations (stacks of dies with different sizes, interposer, . . .);
- Inclusion of the temperature dependency of the material properties;
- Steady state and/or transient regime;
- Computational time needed to obtain the temperature profiles;
- Spatial and temporal resolution of the obtained temperature profiles and of the power maps;
- Use of a discretization of the geometry or mesh-free.

The choice between a physical or a behavioral based model influences most of the characteristics of the model itself and, for this reason, it is further elaborated hereafter. A classification of the models based on this same concept in *white-*, *black-* or *gray-box* approaches has been proposed in [8, 55] and is also reported hereafter.

Physical approaches, also called *white-box* approaches [8], are based on the physical equations governing the phenomenon and on approximation techniques. These methodologies use previously established dynamic equations and the only unknowns are the parameters in the equations. They allow for an easy explanation of the obtained results from a physical point of view, since the underlying laws are explicitly established, but their ability to deal with unknown dynamics and relationships as well as with complex geometries is limited.

Behavioral approaches, also named *black-box* approaches, are, on the other hand, based on the study of the thermal behavior of the analyzed system. They can be based on previously fully simulated data, obtained from an already established modeling technique, such as FEM, or on experimental data. Black-box methodologies give more importance to the response data and their statistical

contents rather than to physical laws underlying the studied phenomenon. If, on the one hand, these approaches are more robust against unknown relationships and give a good behavioral description, they may, on the other hand, lead to physical inconsistency due to over-fitting as well as to unstable models or to results scarcely explainable from a physical point of view [8]. Moreover, since they are based on simulated or experimental data, a time consuming preprocessing phase, which needs to be repeated for every new analyzed structure, is required.

Strategies in between the white- and the black-box methodologies, named *gray-box* approaches, are also possible. They mainly combine previous knowledge and equations together with response data.

In the following, the main FTM methodologies are presented, grouped by the mathematical (or physical) theory underlying them. Some of them calculate the temperature increase θ, over a reference temperature T_{ref}, experienced by the system due to some power dissipation. The value T of the temperature can be obtained afterwards as

$$T = \theta + T_{ref}.$$

Some of the models that will be presented have been originally developed for 2D packaged ICs. However, if the material properties are assumed temperature independent, the extension to 3D-IC structures is normally straightforward and can be performed by applying the superposition principle (cf. Section 1.3.2).

FEM methodologies, which are well established approaches for the thermal analysis of microelectronic devices, are not presented hereafter. This is because the aim of this thesis is to develop a model that can be used to compare the thermal impact of different layout designs and, for this kind of analysis, the running time of the model has to be in the order of seconds or minutes. Nevertheless, since FEM models are accurate and they represent a well established procedure, FEM results will be used as *reference* to validate and compare the methodology presented in the thesis.

1.3.1 Resistance-capacitance network

Classical approach

The firsts CTMs for microelectronic devices have been built approximating the heat paths by means of resistance-capacitance (RC) networks. This approach relies on the electro-thermal analogy that can be established, by coupling temperature difference with voltage difference and current density with heat flux density, between Fourier's law (heat flow by conduction) in the thermal field and Ohm's law (flow of an electrical current) in the electrical field. More precisely, Fourier's law states that, locally,

$$\boldsymbol{q} = -k\boldsymbol{\nabla}T, \tag{1.1}$$

where $\boldsymbol{q}$ is the local heat flux density (W/m^2), k is the material thermal conductivity (W/mK) and $\boldsymbol{\nabla} T$ is the temperature gradient (K/m), while Ohm's law states that, locally,

$$\boldsymbol{j} = -\sigma \boldsymbol{\nabla} V \tag{1.2}$$

where $\boldsymbol{j}$ is the current density (A/m^2), $\boldsymbol{\nabla} V$ is the gradient of the voltage (V/m) and σ is the electrical conductivity of the conductor ($1/\Omega m$). Let's now consider a macroscopic situation in which the current flows through a wire of length Δl and in which all the cross sections of the wire are assumed to be equipotential. By defining the *electrical resistance* between two equipotential surfaces as

$$R_{el} = \frac{\Delta l}{\sigma A}, \tag{1.3}$$

where A is the cross-sectional area of the wire, and by applying some numerical manipulations, equation (1.2) can be rewritten as

$$\Delta V = R_{el} I, \tag{1.4}$$

where I is the current passing through the conductor and ΔV is the voltage difference between the two selected surfaces. This means that the electrical resistance can also be defined in terms of current and voltage difference, i.e. $R_{el} = \frac{\Delta V}{I}$. By exploiting the analogy between equation (1.1) and (1.2), the *thermal resistance* between two points along the heat path can be defined, in the macroscopic scale, as

$$R_{th} = \frac{\Delta T}{Q}, \tag{1.5}$$

where Q is the total heat flux (W) and ΔT is the temperature difference between the two selected points (one is within the conduction region while the other one may be outside the boundary layer).

Unfortunately, although equations (1.1) and (1.2) are analogous, it is erroneous to conclude that there is any practical analogy between thermal and electrical resistances. This is because the difference in conductivity between a conductor and an insulator is of 20 orders of magnitude in the electrical field but of only 3 orders of magnitude in the thermal field. Moreover, the electrical resistance is defined as the difference in potential between two points of a wire divided by the current flowing in the wire between them. The wire cross sections are considered equipotential and outside the wire there is no current flow. Oppositely, true 1D flow of heat can only be approximated because no material comes close to be a perfect insulator and heat radiates through the vacuum. This is why an unambiguous but cumbersome definition is given for the thermal resistance:

Definition 1. The temperature difference between two isothermal surfaces divided by the heat that flows between them is the *thermal resistance* of the materials enclosed between the two isothermal surfaces and the heat flux tube originating and ending on the boundaries of the two isothermal surfaces [50].

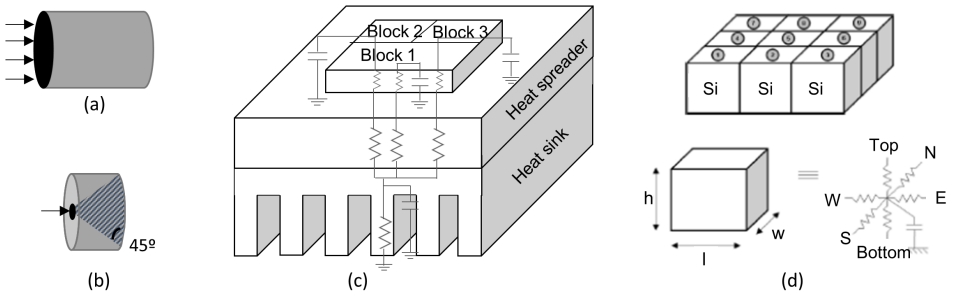


Figure 1.4: (a) Structure for which the thermal resistance is defined when assuming insulating lateral boundary conditions. (b) Illustration of the thermal cone used to compute the spreading thermal resistance with the 45° rule. (c) An example of an RC-network for a packaged IC (from [101]). (d) RC-network with a regular discretization [100].

This situation is really difficult to be obtained in practice but it is the only condition for which equation (1.5) is valid. A really simplistic and unrealistic situation where this assumption is valid can be obtained by modeling a cuboidal or cylindrical structure with uniform power dissipation on one base, heat removal from the opposite side and insulating lateral boundary conditions (Figure 1.4 (a)).

In case equation (1.5) is valid, the thermal resistance R_{th} can be computed based on the geometrical and material properties of the conductor:

$$R_{th} = \frac{\Delta l}{kA} \tag{1.6}$$

where Δl is the distance between the two isothermal surfaces (m), A the area of the isothermal surfaces (m^2) and k the material thermal conductivity (W/mK). This equation has exactly the same structure as equation (1.3) concerning the electrical resistance. It is valid only if the two isothermal surfaces have the same area and if only one material is considered between them. If more than one material is present along the heat path, individual resistances are connected in series or in parallel, depending on the situation. Moreover, mathematical formulas are available in case the two isothermal surfaces have different areas (*spreading* and *constricting* resistance, cf. Section 1.3.2).

The easiest approximation of the thermal spreading resistance (thermal resistance in case the heat source is smaller than the isothermal surface where cooling is applied) is the application of the so called *45° rule*. Underlying this rule is the assumption that the heat spreads out from the source area to the cooling area in a conic/pyramid-shape with a 45° base angle (Figure 1.4 (b)). This is, however, a really poor approximation that may lead to large errors. For this reason, other methodologies based, for example, on the approximation by series expansions of the analytical solution have been developed [52, 63]. This is, however, only

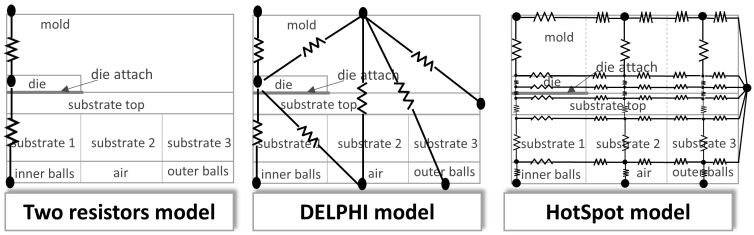


Figure 1.5: Comparison of possible thermal resistance networks for the DELPHI BGA benchmark chip: two resistors model, the DELPHI model and the HotSpot model (adapted from [39]).

possible for simple structures; for more complex geometries an algorithm based on an initial FEM simulation has been recently presented [96].

The impact of boundary conditions (BCs) can also be included in these RC-networks. In case of an isothermal surface, if ΔT in equation (1.5) is considered as the difference between the junction temperature and the temperature of this isothermal surface, by exploiting equation (1.6), the final temperature increase in the location of power dissipation can be obtained. Insulating boundary conditions normally define the heat tube and are intrinsically included in the calculation of the resistances. In case of convection, the corresponding thermal resistance can be computed, starting from Newton's law, as

$$R_{th} = \frac{1}{hA} \tag{1.7}$$

where h is the convection coefficient (W/m^2K).

As already stated, a class of CTM has been built as lumped thermal networks taking advantage of this electro-thermal analogy. This means that various nodes are defined in the system, mainly in the locations where the temperature has to be computed and where the BCs are applied (Figure 1.4 (c)). Extra nodes are normally added to improve accuracy. Node placement is a delicate issue since it may highly affect the final result. More precisely, since the final network should be able to capture the thermal behavior of the whole system, the node placement is case dependent: it depends on the complete packaged structure, the applied cooling solution and the power dissipation. Moreover, if the effect of fine details needs to be included, corresponding nodes have to be added to the model, increasing the overall complexity. The nodes are, then, connected to each other creating a thermal network. Each connection between two nodes, i and j, represents the thermal resistance experienced by the heat while going from i to j. Examples of different resistance networks developed for a particular BGA benchmark chip are shown in Figure 1.5.

These resistance networks are used to compute the steady state solution. The

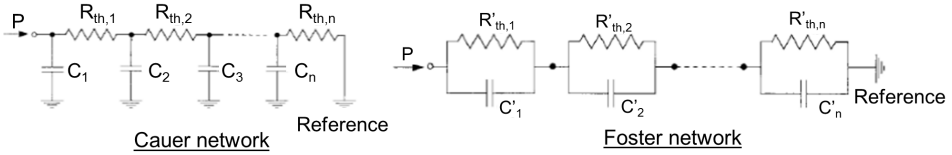


Figure 1.6: Schematic of the Cauer and the Foster RC-networks (adapted from [19]).

addition of *thermal capacitances* allows considering the transient behavior. Thermal capacitances are, indeed, measures of how much heat an object can store and their values indicate how fast the heat is accumulated and/or released from the object. Taking advantage once more of the electro-thermal analogy, the thermal capacitance C is defined as

$$C = Q\frac{\Delta t}{\Delta T}\,, \tag{1.8}$$

which is the ratio between the dissipated power Q and the temperature change over time $\frac{\Delta T}{\Delta t}$. Based on the geometry, the same quantity can be defined as

$$C = c\rho A\Delta l \tag{1.9}$$

where c is the specific heat capacity (J/kgK) and ρ the mass density (kg/m^3).

Based on the electro-thermal analogy, it is possible to define four different kinds of RC-networks, depending on how resistances and capacitances are connected: Cauer I, Cauer II, Foster I and Foster II canonical forms [20]. In thermal analysis, however, just the Cauer I and the Foster I canonical forms of the RC-networks are normally considered (Figure 1.6). This is due, as explained hereafter, to the direct correspondence of these networks to, respectively, *structure functions* and *time-constant representations*. The main characteristics of the two networks are reported hereafter.

- *Cauer network*: each capacitance is grounded. Following this approach, the time needed for the heat to propagate through consecutive sections/materials, each with its own R_{th} and C, is taken into account [35]. This means, in particular, that the effect on the heat sink or on the package of heat dissipation at the junction node is, as it should be, not immediate. Moreover, this kind of network reflects the real, physical setup of the semiconductor. More precisely, in case of layered structures, the values of its R_{th} and C elements can be directly calculated from the material properties and the geometry of the device itself according to equations (1.6) and (1.9). However, as explained earlier in this Section, thermal spreading may be problematic in defining these circuit values [94]. It is important to highlight that this Cauer network can be seen as the representation of the *structure function*, which gives the sum of the thermal capacitances C versus the sum of the thermal resistances R_{th} of the system, measured from the point of excitation towards the ambient [20, 23, 88].

- *Foster network*: each capacitance links two nodes. In this case, individual RC couples do not directly represent the sequence of the physical layers constituting the system (i.e. equations (1.6) and (1.9) do not provide the values of the circuit elements) and the network nodes do not have any physical significance [94]. Moreover, a power dissipation in a certain node results in an immediate temperature increase in the whole system while, in reality, a certain delay is observed due to the capacitances of the materials. However, since the network is the summation of first order responses, it can be easily solved mathematically and the coefficients can be extracted from measured or simulated cooling down curves. These curves are named *thermal impedance curves*, $Z_{th,i}(j,t)$, and they are obtained by dissipating power in point i and monitoring the temperature response in point j (cf. Section 1.3.4) [4, 87, 91]. The network elements, in particular, can be obtained by fitting them with multi-exponential series [35, 94]

$$\frac{T(t) - T_{ref}}{Q} = \sum_{l=1}^{N} R_{th,l}\left(1 - e^{-t/\tau_l}\right) \tag{1.10}$$

  where T_{ref} is an initial reference temperature, $R_{th,l}$ are coefficients representing the resistance values and $\tau_l = R_{th,l}C_l$ are the time constants. This representation allows the creation of a graph, named *time-constant representation*, in which $R_{th}(\tau)$ is plotted as a function of τ. From this graph the elements of the Foster RC-network can be easily read [88, 104].

Transformation algorithms exist to convert the physical meaningful Cauer network into a more easily solvable Foster one and vice versa [33].

Unfortunately, as can be inferred from the definition, thermal resistances are one directional quantities: this means that, in a realistic situation in which the heat spreading is multi-directional, more than one resistance originating from the same node is needed to split the effect into different directions. If the Cauer thermal network is such that the heat flux tubes are clearly defined, then equation (1.6) can be used to compute R_{th} based on the geometrical properties of the system [50]. Otherwise, genetic and optimization algorithms are needed to calibrate the resistance values so that the heat flux is correctly modeled and the overall behavior of the system is captured [19]. In this last case, various steady state FEM simulations with different boundary conditions are run and the values of the components of the network optimized to match the obtained results [101]. The same reasoning is valid for the capacitance values: if the volume each capacitance refers to is clear from the CTM construction, then equation (1.9) can be used to define their values in a Cauer network. Otherwise, optimization algorithms minimizing the error between the CTM and transient FEM results for an exhaustive set of cases, have to be run. This is the case for the classical RC-network modeling approaches, the most famous of which is called *HotSpot* and has been developed by the University of Virginia [38, 101].

The values assigned to the resistance and the capacitance elements in the RC-networks are normally computed assuming a specific value of the ambient temperature and of the temperature rise. This approach is based on the assumption that the thermal system behaves linearly, i.e. that the material properties are not depending on temperature. This simplification is valid if the temperature rise does not exceed $\sim 50°C$. If the temperature variation is higher, the non-linearity effect becomes significant [89]. In [90], the authors propose a methodology to include this non-linearity in the RC-based CTM. The idea is to compute, in a first step, the resistance and the capacitance values, for a fixed network topology, at different temperatures. From these values, the relationship between the network elements and temperature can be extracted and a non-linear CTM is created. This procedure requires, however, to run multiple optimization studies in order to obtain the final, non-linear CTM.

The need of optimization over multiple FEM results can be seen as a drawback of this strategy because this computationally expensive step has to be repeated for each new analyzed structure. Moreover, the temperature is computed just in the locations of the nodes, which, to keep the computational time as low as possible, is normally a relatively small number of points (from less than 10 nodes for DELPHI RC-networks to more than 1000 for HotSpot based models). This means that the temperature map has a low resolution. Despite these drawbacks, once the optimization step has been completed, this methodology can provide a simple and fast estimation of the nodal temperatures under any arbitrary set of BCs [87].

RC-network and finite difference

To overcome some of the issues associated with the classical RC-network approach, RC-networks based on regular discretization of parts of the modeled device have been proposed [38]. Each thermal cube (represented by one node) is connected to its six neighbors with six thermal resistances and it is equipped with a grounded capacitance to account for transient behavior (Figure 1.4 (d)). In this way, the network elements can be computed via formulas (1.6) and (1.9), without any initial FEM simulation. One of the methodologies that works in this sense is named *ICE* [100].

If the finite difference (FD) method is directly applied to solve, over a regular grid, the heat conduction equation

$$\rho(\boldsymbol{x})c(\boldsymbol{x})\frac{\partial T(\boldsymbol{x},t)}{\partial t} = \boldsymbol{\nabla}\cdot[k(\boldsymbol{x},T)\boldsymbol{\nabla}T(\boldsymbol{x},t)] + q(\boldsymbol{x},t), \tag{1.11}$$

where $q(\boldsymbol{x},t)$ is the internal dissipated power density, the coefficients that appear in the $\boldsymbol{G}$ and $\boldsymbol{C}$ matrices resulting from the FD discretization,

$$\boldsymbol{G}T(t) + \boldsymbol{C}\dot{T}(t) = \boldsymbol{p}(t), \tag{1.12}$$

are the same as the corresponding R_{th} and C values derived from the RC-network approach applied to the same grid. Since the solution of this linear system is time consuming for a large number of nodes, only the part of the geometry that is of thermal interest is regularly discretized. The impact of the package is normally included by means of thermal resistances, which are applied as BCs. Moreover, with a regular discretization, the spatial resolution of the temperature profile increases. At the same time, however, the computational time required to obtain the solution does the same, because of the increased dimension of the resulting FD system.

The system of equations (1.12), obtained by applying FD or a regular RC-network discretization, is similar in nature to the one that would be obtained by applying FEM on the same grid, meaning that no significant speed up in computational time is expected. However, the modeled IC considered in [100] includes interlayer channels for liquid cooling. Although, for pure conduction, this CTM based on a regular RC-network is not expected to provide a significant gain in computational time, when the fluid is accounted for, its extension to deal with this new situation is about three orders of magnitude faster than CFD.

1.3.2 Analytical solutions

Full 3D analytical solutions are available only for simplified situations. They mainly deal with the computation of the spreading resistance in different steady state scenarios: single layer structures with one convective BC and HS power dissipation in the center [52], two layers with one convective BC and eccentric heat sources [70, 112] or packaged and layered structure with uniform power dissipation [27]. The mathematical techniques involved in the analytical solution of the heat conduction equation in these cases are mainly separation of variables and Fourier series expansions. In order to apply them, the heat equation is assumed to be linear, meaning that the material properties are temperature independent. This is because this assumption allows the application of the *superposition principle*.

Definition 2. The *superposition principle* states that, for all linear systems, the net response at a given place and time caused by two or more stimuli is the sum of the responses which would have been caused by each stimulus individually.

This means that, if the power Q_A applied in position A generates a temperature increase θ_A and the power Q_B applied in position B generates a temperature increase θ_B, then the simultaneous dissipation of Q_A and Q_B in position A and B respectively generates a temperature increase $\theta = \theta_A + \theta_B$.

In [17] the authors present a semi-analytical solution for the transient regime in case of a multilayer structure with one convective BC and heat generation on multiple layers. Imposing continuity conditions for the temperature and the heat flow across the stack allows to employ an iterative approach to solve for the temperature fields,

which are expressed in Fourier cosine series, on the different layers. The use of an iterative approach means that the method is not fully analytical. Moreover, this solution is obtained for the steady state regime. To solve the transient problem, the *Laplace transform* is used. It transforms the time dependent heat conduction partial differential equation (PDE) and the corresponding BCs into time independent PDE and BCs. Since this new problem is similar in nature to the steady state one, the same approach developed to compute the steady state solution can be applied before transforming back the results to the time domain.

1.3.3 Green's function based modeling approaches

Analytical solutions

The application of Green's function theory to the heat transfer phenomenon is another way to obtain analytical solutions. The main assumption is, again, that the PDE is linear, meaning that the material properties are assumed temperature independent.

Definition 3. A Green's function $G(\boldsymbol{x}, t; \boldsymbol{x}_0, t_0)$ for the heat conduction equation is defined as the solution of the linearized equation, under specific BCs, in position $\boldsymbol{x}$ and time t when an impulsive and localized power is dissipated in position $\boldsymbol{x}_0$ and at time t_0, i.e. if $q(\boldsymbol{x}, t) = \delta(\boldsymbol{x} - \boldsymbol{x}_0)\delta(t - t_0)$ in equation (1.11), where δ is the Dirac delta function.

The function $G(\boldsymbol{x}, t; \boldsymbol{x}_0, t_0)$ depends, of course, on the geometry, materials and BCs of the system but, once it is known, the solution of the general heat conduction equation (1.11) for the temperature increase $\theta(\boldsymbol{x}, t)$ can be computed as

$$\theta(\boldsymbol{x}, t) = \int_\Omega \int_0^t G(\boldsymbol{x}, t; \boldsymbol{x}_0, t_0) q(\boldsymbol{x}_0, t_0) d\boldsymbol{x}_0 dt_0. \tag{1.13}$$

where Ω is the spatial domain in which the thermal problem is defined. The impedance curves $Z_{th,i}(j, t)$, introduced in Section 1.3.1, and the Green's function $G(\boldsymbol{x}, t; \boldsymbol{x}_0, t_0)$ have a lot in common. The main difference is that, while in $Z_{th,i}(j, t)$ the heat source can have a more general shape, it is represented by a step function in time and the temperature response is computed just in point j, in the Green's function approach the heat source is represented by a δ impulse and both the dissipation and the temperature response locations are variable. This basically means that the time derivative of $Z_{th,i}(j, t)$ is equivalent to $G(\boldsymbol{x}, t; \boldsymbol{x}_0, t_0)$ when $\boldsymbol{x} = j$, $\boldsymbol{x}_0 = i$ and $t_0 = 0$. The time derivative is needed to account for the difference between the step power applied to obtain Z_{th} and the impulsive power applied to obtain $G(\boldsymbol{x}, t; \boldsymbol{x}_0, t_0)$. Both a full set of $Z_{th,i}(j, t)$ and of $G(\boldsymbol{x}, t; \boldsymbol{x}_0, t_0)$, covering all the possible heat dissipation and temperature response positions, fully characterizes the system [95] (cf. Section 1.3.4).

Different methodologies have been proposed to compute the values of $G(\boldsymbol{x}, t; \boldsymbol{x}_0, t_0)$ and, in all cases, simplifications and approximations of the geometry have been considered. Various research groups approached the problem analytically using techniques such as separation of variables followed by eigenvalues-eigenfunctions decomposition or cosine series expansion [109, 113], elliptic theta functions [36], symbolic calculator [36], Bessel functions [108], Galerkin integral method [43],.... In most of the cases, the outcomes of these approaches are implicit equations, infinite series and integrals that need to be solved, truncated or numerically approximated. For these reasons, the computational time associated to these methods can be similar in magnitude to the one of FEM simulations [36]. This is especially the case if high resolution is required for the temperature maps. However, one of the main advantages of the FTM approaches based on analytical Green's functions is that they are mesh-free methods. This means that, opposite to what happens for FEM where the whole domain needs to be properly meshed and the solution computed everywhere, the temperature computation can be limited to points of interests [43].

Another aspect that affects computational time is the geometry and BCs dependence of the Green's functions. This means, on the one hand, that a new Green's function needs to be calculated for each new geometry and applied BCs. However, on the other hand, if only the dissipated power density q varies, only the integral in equation (1.13) between the new q and the Green's function needs to be computed. This property drove different authors to propose look-up tables [43, 113] and precharacterization [109] as possible ways to speed up the computations. Moreover, it has to be noted that, because of the boundary effect and the presence of different materials in the modeled device, the $G(\boldsymbol{x}, t; \boldsymbol{x}_0, t_0)$ functions depend on the position where the δ power is dissipated. The response of the system is, indeed, different if the power is dissipated close to a corner of the die stack or in its center.

As mentioned in the beginning of this subsection, analytical solutions cannot be obtained for any geometry and configuration: certain simplifications need to be imposed. In [36], for example, the authors derive a Green's function based transient solution procedure for a single layer cuboidal geometry and nonlinear boundary conditions while in [43] a transient solution is proposed for multilayered structures. Both these approaches are associated with high computational time. In [109] another methodology is presented to solve the latter problem combining the eigen-expansion technique and the electro-thermal analogy. In this way, fully analytical and explicit formulas are provided and, for steady state simulations, a significant reduction in computational time is reported. All these approaches assume that the material properties cannot vary in the horizontal direction and that they are isotropic. In microelectronic devices these approximations are often too strict. There are, indeed, layers where different materials are present (μbumps and underfill, for instance) and, if accounting for the real geometry is too complex and time expensive, at least equivalent orthotropic material properties should be considered.

Semi-analytical solutions

The high required computational time together with the need for a simplified geometry are two of the reasons why a semi-analytical model, which combines FEM and Green's function theory, has been developed. This is the so called *Power Blurring* technique that has been initially developed at the University of California, Santa Cruz. In this approach, the two fundamental ingredients needed to compute the temperature increase in a 3D-IC system are discretized. More precisely, *power maps*, which are matrices storing the information about the dissipated power density on each active layer, and *thermal masks*, which correspond to the Green's functions in the analytical approach, are defined. This means that a grid is needed and, therefore, the methodology is not mesh-free anymore [116].

In this approach, the *thermal masks* are computed by means of FEM, dissipating impulsive power in a small area in the center of each active layer. Each layer is heated up separately and the normalized temperature responses of the system on all the levels of interest are recorded. This means that, if there are N active layers on which we also want to compute the temperature, N^2 thermal masks need to be computed [85]. Moreover, it is important to note that the thermal masks differ from the proper Green's functions because they are restricted to horizontal layers and because the power is dissipated just in one point per layer. This is because the response of the system to HS power dissipation is assumed to be independent of the horizontal position where the HS is dissipated. To take into account the effect of the insulating lateral BCs, the *method of images* is used [37]: one frame of images of the dissipated power map is added all around the original power map (cf. Chapter 3 for more information). Due to the position independence assumption, in steady state $G(\boldsymbol{x};\boldsymbol{x}_0) = G(\boldsymbol{x}-\boldsymbol{x}_0)$ and the integral in equation (1.13) can be seen as a convolution integral

$$\theta(\boldsymbol{x}) = \int_{\Omega} G(\boldsymbol{x}-\boldsymbol{x}_0)q(\boldsymbol{x}_0)d\boldsymbol{x}_0 = (G * q)(\boldsymbol{x}) \tag{1.14}$$

where $*$ is the convolution operator. This has the advantage that fast Fourier transform (FFT) can be implemented to reduce the computational time, allowing higher resolution [86]. Finally, temperature profiles referring to the same die are summed up to obtain the final temperature depending on all the dissipated power in the stack.

The basic geometry for which this method has been developed is really similar to the one for which analytical Green's function solutions are available: a stack of different layers all with the same horizontal dimensions. Cooling is assumed just from the heat sink by applying a convective BC and all the other boundaries are considered adiabatic. The thermal masks are still geometry dependent but their calculation is simplified thanks to the assumption of position independence and to the method of images. Moreover, this semi-analytical approach can deal with any number of stacked layers of different thickness without any significant impact on computational time.

Concerning the transient regime, a combination of convolution in space, based on the Green's function approach, and superposition in time is proposed [84]. The thermal masks are computed for impulsive power dissipation and stored as functions of time. Then, the time dependent temperature responses to the power maps dissipated at each fixed time step are computed separately by convolution. The subsequent superposition in time takes into account the temporal sequence in which the power is dissipated [46].

Various extensions of this methodology have also been presented, mainly for the steady state regime. The first one concerns a pyramid geometry where, for instance, a heat spreader and a heat sink, with much larger footprint areas than the dies, are attached on top of the stack. The impact of the different footprint areas is taken into account by means of extra FEM simulations describing the thermal response of the full system to uniform power dissipations. The error between this solution and the one obtained for the stack configuration, in which just the die stack is modeled, is computed. It is afterwards used as an *error compensation* factor on top of the temperature solution obtained by the *Power Blurring* method in case of the same stack configuration and a general, non-uniform power map [37].

The thermal impact of TSVs on the steady state temperature profiles is considered in [118]. Different sets of thermal masks are computed assigning either silicon or copper material properties to the dies. The thermal masks need, indeed, to be calculated for stacks of uniform material layers. A scan over all the grid elements is performed and, if a certain element is a TSV element, then the thermal mask computed with copper material is used in the convolution, otherwise the one obtained for silicon material. This means that grid element-by-grid element convolution operations are performed, highly increasing the computational time. A low pass filter is then applied to smooth the temperature profiles at the edges between the two areas. Finally, the error compensation, defined, in this case, not only to take into account the difference in shape, but also the difference in materials between the geometries modeled by the FEM and by the *Power Blurring* approach, is applied.

Moreover, parametrization of thermal conductivity, convective heat transfer coefficient and chip thickness are treated in [83] with the aim of building thermal masks without relying on FEM in case of a 2D technology. Finally, two iterative techniques to take into account the dependence of silicon thermal conductivity on temperature are presented in [117] for steady state and transient regime. They are based on selecting proper thermal masks depending on the estimated average or punctual temperature increase in the die.

The *Power Blurring* methodology appears to be fast and to offer basic solutions to almost all the critical points related to the FTM generation. However, it relies a lot on FEM simulations to create thermal masks and error compensation profiles, and just one kind of BC is considered: convection on the top side of the structure and insulation elsewhere. Moreover, it doesn't consider orthotropic material

properties and some of the extensions of the basic methodologies to deal with more complicated situations can be improved.

Another approach that exploits the temperature responses of the system to hot spot power dissipations and the method of images to deal with insulating BCs is presented in [105] for steady state simulations of 3D-ICs. The approach developed in this paper is, however, not based on the Green's function theory but on the superposition principle. This means that superposition of the thermal masks according to the dissipated power density maps is performed, instead of convolution, on each active layer. As a consequence, a much higher computational time and memory usage are needed because the fast Fourier transform cannot be applied.

The chip structure considered in [105] consists of maximum three stacked dies of different thickness (but all with the same horizontal size) separated by back end of line (BEOL) and interface material layers. All the layers have homogeneous material properties and convective BCs are applied on the top and bottom surfaces while insulation is assumed on the lateral sides. A proper function to accurately fit the hot spot temperature responses, obtained by FEM, both in the die where the heat is dissipated and in the other dies has been defined. It depends on five parameters and takes into account the dependency on some geometrical and physical properties: the lateral and vertical dimensions of the dies, the conductivities and the thicknesses of the interface layers, the external thermal resistances, the level on which the HS is applied and the dissipated power. In this way, the final model is independent of FEM simulations

1.3.4 Thermal impedance curves

The *thermal impedance curve* $Z_{th,i}(j,t)$ has already been mentioned multiple times in this literature review. It is basically a 1D-representation of the heat flow; it describes the temporal evolution of the temperature in position j due to power dissipated in point i. When discussing about RC-networks in Section 1.3.1, it was introduced as a way to obtain, by fitting, the values of the resistances and capacitances in a Foster network. In Section 1.3.3, the analogy between the Green's function and the thermal impedance has been underlined. Furthermore, this concept will also appear in the approaches based on model order reduction (Section 1.3.5).

Two points are still open: how to perform the fitting efficiently and how to compute T once the system is characterized by having a full set of $Z_{th,i}(j,t)$ for all the combinations of i and j of interest. Concerning the former one, thermal transient impedance curves can be approximated by multi-exponentials [81]

$$\frac{T(t) - T_{ref}}{Q} = \sum_{l=1}^{N} R_l\left(1 - e^{-t/\tau_l}\right) \tag{1.15}$$

where T_{ref} is an initial reference temperature, R_l are coefficients representing the resistance values in the corresponding Foster RC-network and $\tau_l = R_l C_l$ are the time constants. The difference between *self-impedance*, $Z_{th,i}(i,t)$, and *trans-impedance*, $Z_{th,i}(j,t)$, curves is that, in the former case, the amplitude of the exponentials have to be positive while, in the latter case, it can be negative. In the Laplace domain, this step response corresponds to a system with the following impedance:

$$Z_{th,i}(j,s) = \sum_{l=1}^{N} \frac{R_l}{1+s\tau_l}. \tag{1.16}$$

This is one thermal impedance curve, characterizing the system for power dissipation in point i and temperature computation in location j. For multiple power sources and temperature response locations, the system is fully characterized when the thermal impedance matrix $\boldsymbol{Z}_{th}$ is known. Each element $z_{i,j}(t)$ of this matrix is time dependent and represents the thermal impedance curve $Z_{th,j}(i,t)$. The corresponding time dependent linear system of equations has to be solved to obtain the temperature evolution in the points of interest

$$\boldsymbol{T}(t) = \boldsymbol{Z}_{th}(t)\boldsymbol{Q}(t) + T_{ref}. \tag{1.17}$$

The problem, now, is to find the coefficients in the exponential series describing $Z_{th,i}(j,s)$. Since exponential decays are not orthogonal, the process is sensitive to noise and truncation and, since the multiexponential fitting is a nonlinear minimization problem, it is prone to be trapped in local minima. Three different possibilities have been proposed and, for each of them, pros and cons are presented in [81].

Exploiting the similarities with the Green's function approach, the time dependent temperature increase in point i can also be written as

$$T_i(t) = T_{ref} + \sum_{j=1}^{N} \int_0^t \dot{Z}_{th,j}(i,\tau) Q_j(\tau) d\tau \tag{1.18}$$

where the dot indicates the time derivative and $Q_j(\tau)$ is the power dissipated at time τ in position j. There are two differences between equation (1.13) in Section 1.3.3 and equation (1.18) defined here: the presence, in the last case, of the time derivative and of the dissipated power instead of the dissipated power density. The former discrepancy is due to the difference in the power signal used to characterize the Green's function and the thermal impedance: impulsive in the former case and stepwise in the latter one. The presence in equation (1.13) of q instead of Q is due to the integration in space, which does not occur in equation (1.18).

The thermal impedance curves can also be used in calculating the eigenvalues-eigenfunctions decomposition for the transient problem in a semi-analytical way [34]. The analytical derivation of the formulas for the eigenvalues and

eigenfunctions shows, indeed, that they can be related to the resistances and the capacitances in a corresponding Foster network. However, as the Foster network has no physical meaning and it is specific for each geometry, fittings from FEM simulations or measurements are needed to get the impedance curves Z_{th}. From them, the values of the network elements, the eigenvalues and the eigenvectors can be computed. This method requires a lot of computational effort to be built because the number of needed eigenfunctions and eigenvalues equals the amount of terms considered in the truncated series expansion originated from the separation of variables approach. This number may be higher than hundred thousand. However, when the eigenfunctions and eigenvalues are known for the specific system, the temperature responses to different power dissipation profiles can be quickly computed.

1.3.5 Model order reduction

The idea of *model order reduction* is to describe a complex system, originally defined by n variables, using just k properly selected variables, with $k \ll n$. As a consequence, a certain error is introduced but, at the same time, the solution of the corresponding system of equations is computationally less expensive. In other words, instead of looking for a solution of the original system of equations defined for the variable $\boldsymbol{T}$, where $\boldsymbol{T} \in \mathbb{R}^n$, a new, smaller system is defined and solved for the variable $\boldsymbol{x} \in \mathbb{R}^k$, where $\hat{\boldsymbol{T}} = \boldsymbol{V}\boldsymbol{x}$ is an approximation of $\boldsymbol{T}$ in $\mathbb{R}^k$. The question is now how to compute the projection matrix $\boldsymbol{V}$ and the solution $\boldsymbol{x}$ of the new system. Various methodologies have been presented in literature to optimize the way in which this projection is performed.

The *proper orthogonal decomposition* (POD) methodology [1, 4], also known as Karhunen–Loève transform (KLT) in signal processing, Hotelling transform in multivariate quality control, singular value decomposition (SVD) of $\boldsymbol{T}$ and eigenvalue decomposition (EVD) of $\boldsymbol{T}^T\boldsymbol{T}$ in linear algebra, has as basic ingredient a properly computed set of observations. It basically consists in projecting this set, of possibly correlated data, into a space spanned by a set of orthogonal uncorrelated variables. These variables are ordered and chosen in such a way that the first ones describe the main characteristics of the system and the last ones just small, negligible features. In this way, neglecting the contribution of some of the last variables ensures the error to be minimal. This is the *optimality property* associated to this method.

The POD methodology applied to the thermal modeling of ICs starts by generating the observation matrix, $\boldsymbol{T}_{snap}$, which consists of a collection of N observations, where N is the number of cells where power can be dissipated, from full numerical simulations or measurements. These observations are obtained subsequently heating up each *heating cell* separately and recording the temperature responses in all the locations of interest. Eigenvalues-eigenfunctions decomposition of the

corresponding correlation matrix, $\boldsymbol{C} = \frac{1}{N}\boldsymbol{T}_{snap}\boldsymbol{T}_{snap}^{T}$, is used to project the matrix into a new orthogonal basis. The obtained eigenfunctions, φ_i, are, then, ordered according to the position that the corresponding eigenvalues, λ_i, have in a non-decreasing ordered sequence. If a POD basis of order k is chosen, then, keeping the first k POD modes in this sequence ensures the error to be minimal.

Once the POD modes have been calculated, the N time dependent POD coefficients, $a(t)$, needed to obtain the temperature response to a general power dissipation, have to be computed. They are calculated applying the Galerkin projection method to the discretized heat conduction model in which the POD projected temperature vector, $\hat{\boldsymbol{T}} = \boldsymbol{V}^T\boldsymbol{x}(t)$, is substituted to the full temperature vector $\boldsymbol{T}$ ($\boldsymbol{V}$ is the matrix collecting the first k ordered eigenfunctions).

It is important to note that, to generate $\boldsymbol{T}_{snap}$, all the *heating cells* should be excited independently. This is the same idea as the full characterization of a thermal system by computing the thermal impedance matrix $\boldsymbol{Z}_{th}$. Even though in the original dataset no simultaneous excitations are considered, the obtained POD modes can describe these situations provided that individual excitations are stored. On the other hand, the POD modes are geometry and BCs dependent: their usefulness is limited to the specific system for which they have been derived. This means that, if a new configuration has to be studied, the whole procedure has to be repeated, which normally requires a lot of time. This is mainly due to the construction of the $\boldsymbol{T}_{snap}$ matrix, since multiple full numerical simulations (or measurements) are needed, and also to the eigenvalues-eigenfunctions decomposition, in case the amount of collected data is large (high resolution). However, since the $\boldsymbol{T}_{snap}$ matrix is obtained from full numerical simulations or from measurements, the thermal impact of the package and of small structures (TSVs, BEOLs, μbumps,...) is captured, provided that this effect is included in the FEM and in the considered POD modes. Moreover, once the POD modes have been computed, multiple power maps can be quickly tested since just a new calculation of the POD coefficients is required.

As the dimensions of the reduced order problem are much smaller than the original ones, proper optimization tools or control algorithms can be employed for optimal location, duration and intensity of the power dissipation. However, it is not possible to optimize the system with respect to parameters referring to package, geometry, material and BCs since, for each new configuration, a new $\boldsymbol{T}_{snap}$ matrix is needed.

A slightly different approach consists in building the reduced order model for the steady state regime by means of the *Krylov subspace* method via the *Arnoldi algorithm* [2]. Opposite to the POD strategy, no results from FEM simulations are needed but just the discretization and the system of equations derived from the FEM approach. The direct solution of this linear system, which can be written as $\boldsymbol{GT} = \boldsymbol{p}$, where $\boldsymbol{G}$ and $\boldsymbol{p}$ are known, requires the inversion of the matrix $\boldsymbol{G}$ that, being $\boldsymbol{G}$ a large matrix, is a computationally expensive operation. However, the

Cayley-Hamilton theorem claims that $\boldsymbol{G}^{-1}$ can be obtained as a linear combination of subsequent powers of $\boldsymbol{G}$. This means that $\boldsymbol{T} = \sum_{l=0} \alpha_l \boldsymbol{G}^l \boldsymbol{p}$ and the coefficients α_l are computed minimizing the residuals. The Krylov subspace of order k is, then, defined as $\mathcal{K}_k(\boldsymbol{G}, \boldsymbol{p}) = \text{span}\{\boldsymbol{p}, \boldsymbol{G}\boldsymbol{p}, \boldsymbol{G}^2\boldsymbol{p}, \ldots, \boldsymbol{G}^{k-1}\boldsymbol{p}\}$. The truncation of the linear combination after k terms reduces the complexity of the model, which is now represented in a lower dimensional space, and increases the computational speed. In this approach, therefore, $\boldsymbol{V} = [\boldsymbol{p}|\boldsymbol{G}\boldsymbol{p}|\boldsymbol{G}^2\boldsymbol{p}|\ldots|\boldsymbol{G}^{k-1}\boldsymbol{p}]$ and $\boldsymbol{x} = [\alpha_l]$. It is important to note that, even if no FEM simulations are needed, the Krylov subspace depends on the dissipated power $\boldsymbol{p}$ and, as a consequence, it has to be recomputed for every variation, both in geometry and dissipated power, of the analyzed system.

In the approach presented in [21], the *multivariate moment matching method* is proposed to generate the reduced order model in transient regime. In a first step, the Laplace transform is applied to the system of equations obtained by means of a discretization technique, such as FEM or FD, in order to eliminate the time derivative. Furthermore, from the transformed system, the thermal impedance matrix $\boldsymbol{Z}_{th}(\boldsymbol{s})$ is defined. As already stated in Section 1.3.3, this matrix fully characterizes the system from a thermal point of view, meaning that the vector $\boldsymbol{T}(t)$ can be computed once $\boldsymbol{Z}_{th}(s)$ is known. The idea behind the method is to select the matrix $\boldsymbol{V}$ so that the first moments in the multivariate series expansion of $\boldsymbol{Z}_{th}(\boldsymbol{s}) \in \mathbb{R}^n$, in a chosen linear subspace $\mathbb{R}^k$ around an expansion point $\boldsymbol{\sigma}$, are the same as the ones of its projection $\hat{\boldsymbol{Z}}_{th}(s) \in \mathbb{R}^k$. A *multivariate Taylor* expansion is needed because the variable $\boldsymbol{s}$ represents both the angular frequency and the information concerning the BCs. The projecting matrix $\boldsymbol{V}$, which is computed from the moments of the reduced impedance matrix $\hat{\boldsymbol{Z}}_{th}(\boldsymbol{s})$, can be obtained using a methodology derived from the Krylov subspace theory. Once $\hat{\boldsymbol{Z}}_{th}(s)$ and the projected system of equations are known, the vectors $\boldsymbol{x}(t)$ and $\hat{\boldsymbol{T}}(t)$ can also be calculated. This approach becomes computationally expensive if multiple heat sources are present. For this reason, in [22] the authors proposed an extension to consider these situations. It is actually based on the partitioning of the system into multiple subregions, each containing few heat sources. The thermal problem is solved in each of them separately and the results are combined afterwards.

1.3.6 Multi-scale and multi-grid approaches

The thermal management of packaged microprocessors involves several decades of length scales: the heat is generated in the transistors, which dimensions are in the *nm* order of magnitude and transferred to the ambient through heat sinks, which are in the order of *cm*. For this reason multi-scale and multi-grid approaches have been proposed [4].

The main goal of a *multi-scale* approach is to combine the detailed information about the part of interest in the modeled structure (the die stack in our case), which is usually small with respect to the whole system, with the coarser information

about the large part around it (package and system). This last part can, indeed, be modeled with lower accuracy since just its general thermal behavior is sufficient to obtain accurate enough results. Modeling the complete structure in a detailed way would require too high computational effort. Different modeling methodologies can be combined to deal with the detailed and the coarse part of the model.

The multi-scale approach proposed in [4] aims to reduce the computational time of transient finite element simulations. The main idea is to initially study the package transient thermal response to a uniform heat source applied in the chip position, which is, at this stage, considered as a solid block with internal material homogeneity and power uniformity. The transient heat fluxes and temperature distributions on the top and the bottom of the chip are computed. They are used, in a second stage, as BCs for the FEM simulation of the detailed chip, which includes finer details and material heterogeneity. The real, non-uniform, transient power map is now applied to the chip and the temperature responses at different time steps are recorded.

Multi-grid methods implement a similar idea. As the name suggests, they are grid based methodologies that efficiently solve the linear system, obtained by applying the FD discretization to the heat transfer problem, on different grid levels. More precisely, they solve the high frequency parts of the problem on finer grids and the lower frequency ones on coarser grids. A geometric multi-grid approach has been proposed in [54]. The grid used in the FD approximation of the heat conduction PDE is subsequently coarsened, from grid size $\bar{h}$ to $2\bar{h}$. At each step, few relaxation or smoothing steps are applied to remove high frequency error. Then, by means of a restriction operator, the residual problem is transferred to a coarser grid. It is, then, solved again applying few smoothing steps to remove lower frequency error, before another coarsening is applied. When the coarser level is reached, the solution is mapped back to the original finer grid by means of interpolating operators. The time complexity of this methodology results to be linear in the number of unknowns, with a small constant factor.

1.3.7 Summary table

In the following Table, a compact overview of the different approaches that have been extensively described in this Section is provided. In particular, the models are compared according to the distinctive and desirable properties of a FTM for 3D-ICs, which are the same listed at the beginning of Section 1.3. The Table provides, therefore, an easy way to compare the pros and cons of each presented methodology.

Method	Category	White- / gray-/ black-box	Need of FEM or measure-ments	Multiple layers/ Het-erogeneous materials	μstructures	Pyramidal structure/ other geometries	Coefficient tempera-ture depen-dence	Transient	Computa-tional time	Temperature profiles/ power maps resolution	Mesh/ Mesh free
RC classical network [4, 19, 35, 52, 63, 101]	RC Network	Gray	Yes, full and multiple FEM	Limited	No	Yes	Yes, com-bination of multiple RC-networks	Yes	Low after comp. expensive pre-processing	Low	Few nodes
RC regular network [100]	RC Network	White	No, just FD system of equations	Yes	Yes, fine grid needed	Can be ex-tended	Yes, coefficients updated at each time step	Yes	Middle-High	Middle	Partially meshed
Analytical solutions for spreading resistance [27, 52, 70, 112]	Analytical	White	No	Limited	No	No	No	No	Low	High	Mesh free
Semi-analytical solutions for multilayer structures [17]	Analytical	White	No	Yes, but just silicon	No	No	No	Yes	Middle	High	Mesh free
Analytical Green's function [36, 43, 109, 113]	Green's function, analytical	White	No	Limited	No	No	No	Yes	Middle	High	Mesh free
Semi-analytical Green's function [37, 46, 83–86, 116–118]	Green's function, semi-analytical	Gray	Yes, simplified FEM simulations	Yes	Yes	Yes, via cor-rection	Yes, iterative	Yes	Middle	High	Partially meshed

Method	Category	White- / gray- / black-box	Need of FEM or measurements	Multiple layers/ Heterogeneous materials	μstructures	Pyramidal structure/ other geometries	Coefficient temperature dependence	Transient	Computational time	Temperature profiles/ power maps resolution	Mesh/ Mesh free
Superposition [105]	Superposition, semi-analytical	Gray	No, modeling limited to a set of parameters	Max 3 layers	No	No	No	No	Middle-High	High	Partially meshed
Impedance curves	Impedance curves	Black	Yes, full and multiple FEM	Yes	Yes	Yes	Partially, combined effect of multiple source not included	Yes	Middle-High	Low-Middle	Few nodes
Proper orthogonal decomposition [1,4]	Model order reduction	Black	Yes, full and multiple FEM	Yes	Yes	Yes	Partially, combined effect of multiple source not included	Yes	High, low for different power but same structure	High	Mesh
Krylov subspace [2]	Model order reduction	White	No, just FEM system of equations	Yes	Yes	Yes	No	No	Middle	High	Mesh
Multivariate moment matching method [21,22]	Model order reduction	White	No, just FD/FEM system of equations	Yes	Yes	Yes	No	Yes	Middle	Low	Mesh
Multi-grid [54]	Multi-grid	White	No, just FEM system of equations	Yes, each layer at least two levels of elements	No	Yes	Yes	Yes	Low-Middle	High	Mesh

1.4 Goals of this work

The previous discussion showed that multiple approaches have been proposed to efficiently and accurately model the thermal behavior of ICs/3D-ICs. The main goal of this doctoral research project is to develop a user-friendly thermal modeling methodology for 3D-ICs that can be adopted during the design phase to thermally improve/optimize the system. This means that it should be feasible to obtain accurate enough temperature maps in all the dies of the 3D package for different values of the *design parameters* in the time range of minutes or seconds.

More precisely, the developed CTM/FTM should:

- be able to deal with both steady state and transient regime;
- be easy to use;
- be able to quickly test and compare the thermal impact of different design parameters;
- be considerably faster, both in execution and creation, than the classical FEM approach. Concerning the run time, the goal is to develop a methodology that is at least two orders of magnitude faster than FEM;
- be applicable to 3D-ICs real cases, meaning that it should be able to include the local and global thermal impact of structures proper of the 3D-IC stacking technology (e.g. μbumps and TSVs);
- be able to include the thermal impact of the package used in each specific situation. In this thesis, just packages with air-convection cooling are considered; more advanced cooling solutions for high power applications, as liquid cooling or jet impingement, are not considered;
- be able to accurately reproduce the results obtained by full, detailed and validated FEM simulations. Accuracy is required both on peak temperatures and on the whole active layers. For this reason, both the relative error on high temperatures and the average error with respect to FEM are expected to be below 5% in the steady state regime. In the transient regime, a metric based on the absolute error is considered and, in this frame, the error is expected to be below 2-3°C;
- result in temperature fields that are highly resolved both in space and time. In this way, both the peak temperatures and the spatial and temporal temperature variations on the active regions are captures;
- account for the temperature dependency of the material properties;
- be usable to optimize and to compare different 3D-ICs designs and floor plans;

- avoid to compute the temperature maps in regions that are of low thermal interest;
- give the possibility to be run as a "black box" in such a way that proprietary information (dimensions, materials, ...) do not have to be disclosed if, for example, just the power dissipation design has to be optimized.

As already mentioned, the user should be able to quickly change the values of some *design parameters* to test their thermal impact. They can actually be grouped in two sets, depending if they are related to the stack itself (geometry and materials) or to the BCs. More precisely, these parameters are:

- related to the stack:
  - number of the dies;
  - thickness of the dies;
  - number of layers with high thermal resistance (BEOL, interface,. . .);
  - thickness of the resistive layer;
  - material of the resistive layers;
  - horizontal dimension of the stack;
  - number and layout of the μbumps/TSVs;
- related to the BCs:
  - type of package (plastic or metal materials);
  - type of applied cooling solution (passive or active cooling);
  - location, intensity and time variation of the dissipated power.

Moreover, the possibility of the FTM to be extended to geometries closely related to the 3D-IC stacking one (as interposer (Section 8.2) or stack of dies with different sizes (Section 8.3)), is also a desired characteristic. The performance of the developed FTM, with respect to the aforementioned success criteria, is evaluated by comparing it to the corresponding FEM model for an analogous structure. The FEM models are, therefore, treated as *reference solutions* against which the FTM is validated and evaluated.

Based on these requirements and on the analysis of the pros and cons of the various approaches presented in literature (cf. summary Table in Section 1.3.7), a semi-analytical Green's function methodology has been chosen as the base on which the model presented in this thesis has been built. First of all, this is a gray-box approach and, thus, it combines and mitigates the pros and cons of the more extreme black- and white-box methods. It allows for high temperature resolution profiles and, even if FEM models are required to compute the Green's

functions, their number and complexity is much lower than in case of other kinds of behavioral methods. Moreover, possible solutions to include the package and the microstructures thermal impact have already been proposed in literature. However, there are still some points that can be improved and the work presented in this thesis goes in this direction.

1.5 Original contributions of this work

As stated in the previous Section, the research work of this doctoral thesis is based on the semi-analytical Green's function approach, originally presented as *Power Blurring* in [37, 46, 83–86, 116–118], and on the *superposition work* published in [105]. We have further elaborated on this topic by improving and extending what the original authors of the *Power Blurring* methodology published in literature.

More precisely, the main contributions of this thesis are listed hereafter.

Thermal impact of μbumps The vertical integration between different dies in the 3D-IC stack is performed by means of metallic interconnections (μbumps), which are good thermal conductors, surrounded by underfill material, which has low thermal conductivity. The number, the placement and the layout of the μbumps affects the vertical heat transfer between the chips and, therefore, it may have a strong impact on the final temperature profile. A methodology able to include the thermal impact of μbumps in the steady state regime has been developed and is presented in Chapter 4. Differently to what has been proposed for TSVs in *Power Blurring*, in our approach the scan over the nature of the grid elements is not needed. As a consequence, the convolution can be performed on a matrix base and not element by element, which has a positive impact on the computational time. The algorithm has been developed for stacks of two dies in stack configuration, meaning that only the die stack is actually modeled and the thermal impact of the package is included by means of convective BCs with constant coefficients. The comparison between the results obtained by the FTM and by analogous FEM simulations (no package and comparable resolution) demonstrates a percentage error on the temperature increase lower than 2% combined with a 20 times improvement in runtime. The local and global thermal impact of μbumps was not included in the published *Power Blurring* technique.

Transient regime The *Power Blurring* approach deals with the transient regime by means of 2D-convolution in space and superposition in time. In Section 2.5.3, a way to obtain transient temperature maps using 3D-convolution (two spatial dimensions and one time dimension) is presented. This technique speeds up the calculations without affecting accuracy. The exact speed up with respect to FEM depends on the spatial resolution, on the amount of

considered time steps and on the temporal variation of the applied power maps.

Package impact in transient regime Convective BCs with constant convective coefficients h are normally applied on top and bottom of the stack configuration. By properly selecting the value of h, the equivalent resistive impact of the package, which is not modeled in the stack configuration, can be included. However, the capacitive property of the non-modeled sections of the geometry cannot be considered by just applying appropriate BCs. This means that the speed at which the device heats up or cools down is not correctly modeled in the stack configuration. This happens on top of another phenomenon, present also at steady state, that is not included in the model for the stack geometry: the thermal spreading due to the larger footprint area of the package than of the die stack. In *Power Blurring* a methodology to include the package thermal spreading has been proposed for steady state in case of pyramid geometries. In this work, we extended this procedure to transient regime and to more general package configurations. For the case study presented in Chapter 5, the solution is obtained more than two orders of magnitude faster than with analogous FEM and the absolute error on the maximum temperature increase is lower than 3°C.

Transient temperature in selected points The *point-FTM* allows to compute the transient evolution of the temperature profiles, including the package thermal impact, just in few selected points with the same accuracy as the fully resolved model but more quickly (cf. Section 5.4.6).

Temperature dependency of material properties The value of the silicon thermal conductivity k_{Si} depends, with a negative feedback loop, on temperature. The approaches based on Green's functions assume the heat conduction equation to be linear, neglecting, therefore, the temperature dependency of the material properties. In the published research, an iterative method has been presented to include the temperature dependence of silicon in steady state and transient regime. In Chapter 6 we propose to use the Kirchhoff transformation to this aim, which is a one time transformation that does not need iterations and, therefore, it does not directly affect computational time.

Different geometries The *Power Blurring* method has been developed and validated considering a particular package structure with heat removal just through one boundary. In this thesis, we present different package configurations for 3D-ICs stacks as well as other geometries strictly related to the 3D technology: interposer (Section 8.2) and stack of dies with different footprint sizes (Section 8.3).

Case studies The developed methodology has also been applied to different case studies illustrating its easiness-to-use and applicability: analysis of the impact of die thinning on the temperature profiles (Section 9.2), transient analysis considering load switching between different cores (Section 9.3.1),

thermal comparison between a 2D and a 3D configuration (Section 9.3.2), thermal impact of different underfill material (Section 9.4.1) and analysis of the maximum temperature as a function of the μbumps amount and layout (Section 9.4.2).

Other minor contributions of this work are:

Computation of thermal masks In the *Power Blurring* approach, the thermal masks are computed from a 3D FEM, dissipating hot spot power in the center of each active layer. In our approach a 2D, semi-infinite, axisymmetric FEM model is used to compute them. This allows the reduction of the time needed both to create and to solve the FEM model, as well as the complete elimination of the thermal effect of the lateral boundaries (Section 2.5.1).

Number of images In the *Power Blurring* approach, a single frame of images is considered all around the original power map to model the insulating, lateral BCs. However, in theory, an infinite number of images is needed to model insulation. In this work, we presented an algorithm to define how many images are actually needed to simulate this condition with a certain accuracy (Chapter 3).

Computation of uniform temperature In the proposed FTM, all the thermal information concerning the stack configuration of the system is stored in the thermal masks. An extremely fast way to compute the temperature increase due to uniform power dissipation, starting from the information stored in the thermal masks, is proposed for both steady state and transient regime (Sections 3.4.1 and 3.6).

Orthotropic material properties In the basic *Power Blurring* approach, one single isotropic material is allowed per layer. However, in 3D technology multiple materials may be present on the same horizontal layer, even if the package is not considered: μbumps and underfill, TSVs and silicon, ... For this reason, if a complex model including small details wants to be avoided, at least equivalent, homogenized properties should be used for these layers. However, since the small, embedded structures are typically longer in one direction, homogenized material properties are orthotropic. The model presented in this thesis accounts for this characteristic (Section 2.4).

Package impact in steady state regime While dealing, in steady state, with geometries different from the stack configuration, the *Power Blurring* approach proposes an error compensation based on the thermal responses of the stack and package configuration to uniform power dissipation. The authors applied this methodology to a system with a heat spreader and a heat sink on top of the die stack. In this thesis, the approach is extended to different package configurations. Moreover, in order to speed up the procedure, a multi-level point of view has been adopted, allowing considerable simplifications of the underlying FEM model (Section 5.3).

1.6 Outline

This thesis is divided into three parts. **Part I**, entitled *"Model for the Stack Configuration"*, presents the fundamental methodology on which the developed FTM is based. Chapter 1 is, in particular, dedicated to a comprehensive literature review while Chapter 2 to a clear description of the Green's function theory, of the convolution based method and of its limitations. The method of images is presented as a way to cope with the finite dimensions of real devices in Chapter 3. The method presented in this Part is able to provide satisfactory temperature estimations for geometries (named die stacks) constituted by multiple layers stacked on top of each other. All the layers need to have the same horizontal size and each single layer has to be made of a homogeneous material with temperature independent properties.

In **Part II** of this thesis, entitled *"Overcoming Limitations of the Stack Model"*, possible solutions to close the gap between the structures that can be analyzed by the model for the stack configurations presented in Part I and the real 3D-ICs are proposed. In Chapter 4, in particular, a methodology is presented to deal with the thermal impact of specific μbump arrays layouts in case of two dies stacks, in face-to-face configuration and in the steady state regime. Chapter 5 describes a multi-level approach to account for the spreading and capacitive effect of the package in which the die stack is included, both in the steady state and in the transient regime. The Kirchhoff transformation is presented in Chapter 6 to deal with the temperature dependency of silicon thermal conductivity in real devices. This procedure, which is not iterative, is fully developed for the steady state regime, while some indications are given for the transient regime.

In **Part III** of this thesis, entitled *"Experimental Validation & Case Studies"*, the model is successfully validated with respect to experimental results both in the steady state and in the transient regime (Chapter 7). Moreover, in Chapter 8, it is proved that approaches, similar to the one presented to deal with the package thermal spreading, can be implemented to apply the developed fast thermal model to different geometries commonly available in microelectronic applications (side by side integration on an interposer and stack of dies with different sizes). In Chapter 9, various case studies, related to realistic analyses that may be needed during the design phase of an IC, are presented showing the applicability and the easiness-of-use of the model.

Finally, in Chapter 10, a summary of the research presented in this thesis is provided together with the general conclusions and some suggestions for future work.

Chapter 2

Convolution Based FTM for Infinite Structures

2.1 Introduction

In this Chapter, the methodology that forms the basis of this convolution based FTM is presented. First, the theoretical background concerning the Green's function theory is described together with the main assumptions that are needed to apply this methodology. The main constraints, due to these assumptions, on the thermal modeling of real, air cooled, 3D-ICs (as the ones described in Section 1.2.2 and illustrated in Figure 1.3) are listed in Section 2.4. One of them is the necessity to consider infinitely large structures and this is the reason why this Chapter deals with *infinite structures*. In order to overcome these limitations, appropriate corrections, which will be described in the following Chapters, have been developed. Figure 2.1 shows both the geometry for which the algorithm presented in this Chapter is valid, and the basic blocks of the developed algorithm. It consists, in particular, in the convolution between structure dependent temperature responses to hot spot power dissipation (cf. Section 2.5.1) and user defined power maps (cf. Section 2.5.2). The procedure to obtain these two fundamental ingredients, their characteristics and the algorithm to calculate the steady state and transient temperature profiles are presented in details in this Chapter.

2.2 Theoretical background

The FTM methodology that has been developed in this thesis is based on Green's functions.

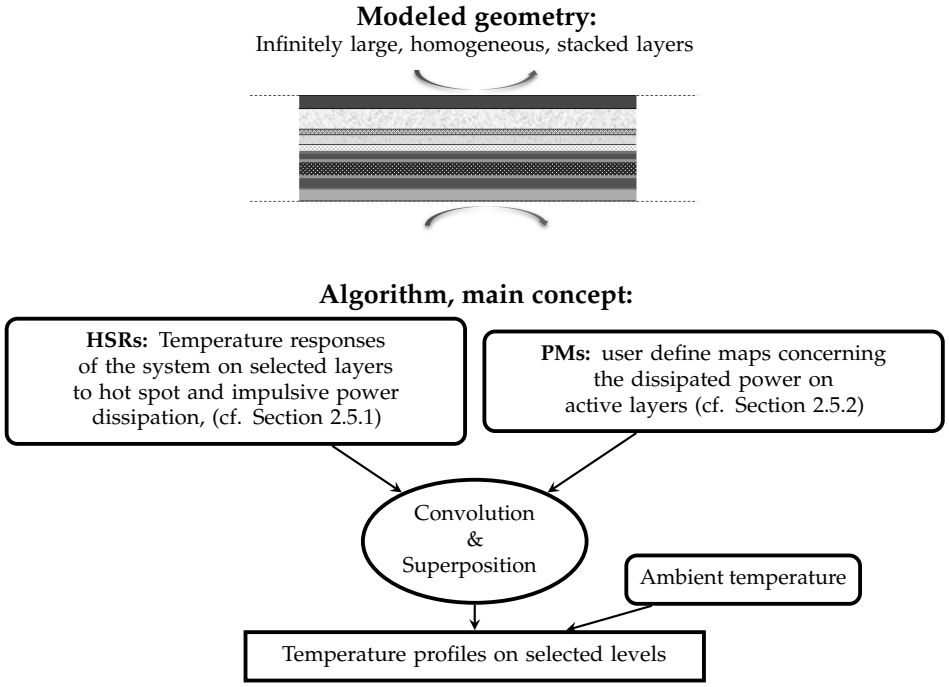


Figure 2.1: Modeled geometry and main concept of the algorithm described in Chapter 2.

Definition 4. A *Green's function*, $G(\boldsymbol{\xi};\boldsymbol{\xi}_0)$, of a linear differential operator $L = L(\boldsymbol{\xi})$ acting on a function at a point $\boldsymbol{\xi}_0$ in a subspace Ω of the Euclidean space $\mathbb{R}^n$, is any solution of

$$LG(\boldsymbol{\xi};\boldsymbol{\xi}_0) = \delta(\boldsymbol{\xi} - \boldsymbol{\xi}_0) \tag{2.1}$$

where $\delta(\boldsymbol{\xi} - \boldsymbol{\xi}_0)$ is the Dirac delta.

The importance of Green's functions is that they can be exploited to solve PDEs of the form

$$Lu(\boldsymbol{\xi}) = f(\boldsymbol{\xi}) \tag{2.2}$$

where $f(\boldsymbol{\xi})$ is a given function.

If a function G satisfying equation 2.1 can be found for the operator L, then, multiplying equation (2.1) by the given function $f(\boldsymbol{\xi}_0)$ and integrating over Ω with respect to the $\boldsymbol{\xi}_0$ variable, the following result is obtained:

$$\int_\Omega LG(\boldsymbol{\xi};\boldsymbol{\xi}_0) f(\boldsymbol{\xi}_0) d\boldsymbol{\xi}_0 = \int_\Omega \delta(\boldsymbol{\xi} - \boldsymbol{\xi}_0) f(\boldsymbol{\xi}_0) d\boldsymbol{\xi}_0 = f(\boldsymbol{\xi}).$$

From equation (2.2)

$$\int LG(\boldsymbol{\xi};\boldsymbol{\xi}_0)f(\boldsymbol{\xi}_0)d\boldsymbol{\xi}_0 = f(\boldsymbol{\xi}) = Lu(\boldsymbol{\xi})$$

and, since L is linear and acts only on $\boldsymbol{\xi}$, it can be taken outside the integral (which is over $\boldsymbol{\xi}_0$) to have

$$Lu(\boldsymbol{\xi}) = L\int G(\boldsymbol{\xi};\boldsymbol{\xi}_0)f(\boldsymbol{\xi}_0)d\boldsymbol{\xi}_0 \Rightarrow u(\boldsymbol{\xi}) = \int G(\boldsymbol{\xi};\boldsymbol{\xi}_0)f(\boldsymbol{\xi}_0)d\boldsymbol{\xi}_0 + a(\boldsymbol{\xi}). \tag{2.3}$$

where $a(\boldsymbol{\xi})$ is the solution of the associated homogeneous equation $Lu(\boldsymbol{\xi}) = 0$ and is determined by the BCs. For infinite space, for example, $a(\boldsymbol{\xi}) = 0$. This means that, if the Green's function G is obtained for a certain differential operator L, then the solution of the PDE is known for any input data $f(\boldsymbol{\xi})$, at least in an integral form. G is the impulse response of the system to a δ input and it can also be called *fundamental solution associated to L*.

When the operator L is *translation invariant*, i.e. it has constant coefficients with respect to $\boldsymbol{\xi}$, and it acts over $\Omega = \mathbb{R}^n$, the Green's function depends only on the distance $\boldsymbol{\xi} - \boldsymbol{\xi}_0$ between the location $\boldsymbol{\xi}_0$ of the δ perturbation and the position $\boldsymbol{\xi}$ where the system response is considered. Under this condition, therefore,

$$G(\boldsymbol{\xi};\boldsymbol{\xi}_0) = G(\boldsymbol{\xi} - \boldsymbol{\xi}_0) \tag{2.4}$$

and the solution of $Lu(\boldsymbol{\xi}) = f(\boldsymbol{\xi})$ can be written as a convolution operator:

$$u(\boldsymbol{\xi}) = \int G(\boldsymbol{\xi} - \boldsymbol{\xi}_0)f(\boldsymbol{\xi}_0)d\boldsymbol{\xi}_0 \tag{2.5}$$

($a(\boldsymbol{\xi}) = 0$ because $\Omega = \mathbb{R}^n$). On top of the advantage of having a Green's function depending on just one variable, the translation invariance property of L allows the application of the *convolution theorem*.

Theorem 1 (Convolution theorem)**.** *The Fourier transform translates between convolution and multiplication of functions. If $f(\boldsymbol{\xi})$ and $g(\boldsymbol{\xi})$ are integrable functions with $\hat{f}(\hat{\boldsymbol{\xi}}) = \mathcal{F}\left\{f(\boldsymbol{\xi})\right\}$ and $\hat{g}(\hat{\boldsymbol{\xi}}) = \mathcal{F}\left\{g(\boldsymbol{\xi})\right\}$ as Fourier transforms, and if*

$$p(\boldsymbol{\xi}) = (f * g)(\boldsymbol{\xi}) = \int f(\boldsymbol{\xi}_0)g(\boldsymbol{\xi} - \boldsymbol{\xi}_0)d\boldsymbol{\xi}_0$$

then

$$\hat{p}(\hat{\boldsymbol{\xi}}) = \hat{f}(\hat{\boldsymbol{\xi}})\hat{g}(\hat{\boldsymbol{\xi}}).$$

Under suitable conditions the Fourier transform is invertible and $p(\boldsymbol{\xi})$ can be retrieved:

$$p(\boldsymbol{\xi}) = \mathcal{F}^{-1}\left\{\hat{p}(\hat{\boldsymbol{\xi}})\right\}.$$

2.2.1 Superposition vs convolution

It is important to note that, since $G(\boldsymbol{\xi} - \boldsymbol{\xi}_0)$ is symmetric with respect to $\boldsymbol{\xi}_0$, convolution is physically the same as *superposition*, just the way to compute it changes. Let's assume that the solution $a(\boldsymbol{\xi})$ of the associated homogeneous equation is zero. If this is not the case, the value of $a(\boldsymbol{\xi})$ can be added afterwards to the superposition/convolution results obtained assuming $a(\boldsymbol{\xi}) = 0$. According to the superposition principle

$$\left.\begin{aligned} &Lu(\boldsymbol{\xi}) = f(\boldsymbol{\xi}) \\ &f(\boldsymbol{\xi}) = \sum_{i=1}^{n} f_i(\boldsymbol{\xi}) \\ &Lu_i(\boldsymbol{\xi}) = f_i(\boldsymbol{\xi}) \quad i = 1, \dots, n \end{aligned}\right\} \Rightarrow u = \sum_{i=1}^{n} u_i(\boldsymbol{\xi}). \tag{2.6}$$

To prove the equivalence between convolution and superposition, let's define a partition $\{\Omega_i\}$ of Ω, i.e. $\bigcup_i \Omega_i = \Omega$ and $\Omega_i \cap \Omega_j = \emptyset$ for $i \neq j$. Let's now define the system response $\tilde{u}_i(\boldsymbol{\xi})$ to *unit* perturbation $Q_{\Omega_i}(\boldsymbol{\xi})$ in Ω_i as the solution of

$$L\tilde{u}_i(\boldsymbol{\xi}) = Q_{\Omega_i}(\boldsymbol{\xi}) \tag{2.7}$$

where

$$Q_{\Omega_i}(\boldsymbol{\xi}) = \begin{cases} 0, & \text{if } \boldsymbol{\xi} \notin \Omega_i, \\ \frac{1}{|\Omega|}, & \text{if } \boldsymbol{\xi} \in \Omega_i. \end{cases} \tag{2.8}$$

It is now clear from the definitions of $\tilde{u}_i(\boldsymbol{\xi})$ and $Q_{\Omega_i}(\boldsymbol{\xi})$ that

$$\lim_{\Omega_i \to \boldsymbol{\xi}_i} Q_{\Omega_i}(\boldsymbol{\xi}) \overset{(2.8)}{=} \delta(\boldsymbol{\xi} - \boldsymbol{\xi}_i) \tag{2.9}$$

and

$$\lim_{\Omega_i \to \boldsymbol{\xi}_i} \tilde{u}_i(\boldsymbol{\xi}) \overset{(2.9)}{=} G(\boldsymbol{\xi}; \boldsymbol{\xi}_i). \tag{2.10}$$

The perturbation function $f(\boldsymbol{\xi}) = \sum_{i=1}^{n} f_i(\boldsymbol{\xi})$ can be approximated as the sum of stepwise functions with disjoint supports in Ω_i:

$$f_i(\boldsymbol{\xi}) \approx \begin{cases} 0, & \text{if } \boldsymbol{\xi} \notin \Omega_i \\ \frac{\int_{\Omega_i} f_i(\boldsymbol{\xi}) d\boldsymbol{\xi}}{|\Omega_i|} := \frac{f_i}{|\Omega_i|}, & \text{if } \boldsymbol{\xi} \in \Omega_i \end{cases} \overset{(2.8)}{\Longrightarrow} \begin{matrix} f_i(\boldsymbol{\xi}) \approx f_i Q_{\Omega_i}(\boldsymbol{\xi}) \\ \lim_{\Omega_i \to \boldsymbol{\xi}_i} f_i = f(\boldsymbol{\xi}_i) \end{matrix}. \tag{2.11}$$

Due to linearity of the differential operator L

$$Lu_i(\boldsymbol{\xi}) = f_i Q_{\Omega_i}(\boldsymbol{\xi}) \overset{(2.7)}{\Rightarrow} u_i(\boldsymbol{\xi}) = f_i \tilde{u}_i(\boldsymbol{\xi}). \tag{2.12}$$

Considering the $u_i(\boldsymbol{\xi})$ functions as defined in (2.12) and taking the limit of the discretization to points,

$$\begin{aligned} u(\boldsymbol{\xi}) &\overset{(2.6)}{=} \lim_{\Omega_i \to \boldsymbol{\xi}_i} \sum u_i(\boldsymbol{\xi}) \overset{(2.12)}{=} \lim_{\Omega_i \to \boldsymbol{\xi}_i} \sum \tilde{u}_i(\boldsymbol{\xi}) f_i \\ &\overset{(2.10),(2.11)}{=} \int_\Omega f(\boldsymbol{\xi}_i) G(\boldsymbol{\xi}; \boldsymbol{\xi}_i) d\boldsymbol{\xi}_i = \int_\Omega f(\boldsymbol{\xi}_i) G(\boldsymbol{\xi} - \boldsymbol{\xi}_i) d\boldsymbol{\xi}_i \end{aligned} \tag{2.13}$$

where the last equality holds due to the translation invariance property of $G(\boldsymbol{\xi}; \boldsymbol{\xi}_i)$. This proved that, under certain conditions, convolution is just a different way to apply superposition. In Section 2.5.3 it will be proved that convolution is computationally much faster than superposition.

2.2.2 Green's functions for thermal modeling of 3D-ICs

As already introduced in Section 1.3.1, the physics equation underlying the heat conduction phenomenon is

$$\rho(\boldsymbol{x}) c(\boldsymbol{x}) \frac{\partial T}{\partial t}(\boldsymbol{x}, t) = \boldsymbol{\nabla} \cdot [k(\boldsymbol{x}, T) \boldsymbol{\nabla} T(\boldsymbol{x}, t)] + q(\boldsymbol{x}, t). \tag{2.14}$$

The differential operator L for this equation is, therefore,

$$L = \rho(\boldsymbol{x}) c(\boldsymbol{x}) \frac{\partial}{\partial t} - \boldsymbol{\nabla} \cdot [k(\boldsymbol{x}, T) \boldsymbol{\nabla}] \tag{2.15}$$

that is not linear because the thermal conductivity is temperature dependent and not translation invariant because, in general, the material properties depend on the position $\boldsymbol{x}$. This means that, to be able to apply the convolution strategy based on Green's function theory we need to assume that

1. the thermal conductivity is temperature independent;
2. the material properties are position independent.

Concerning the first point, the temperature dependency is initially neglected, assuming $k(\boldsymbol{x}, T) = k(\boldsymbol{x}, \bar{T})$, where $\bar{T}$ is the expected average temperature value during chip operation. The simplification introduced by the second assumption is, however, not acceptable for 3D-ICs: multiple materials are, indeed, used in the real devices. A less strict assumption is that the material properties depend just on the vertical direction, i.e. $k(\boldsymbol{x}) = k(z)$, $\rho(\boldsymbol{x}) = \rho(z)$ and $c(\boldsymbol{x}) = c(z)$ and

$$L = \rho(z) c(z) \frac{\partial}{\partial t} - \boldsymbol{\nabla} \cdot [k(z) \boldsymbol{\nabla}]. \tag{2.16}$$

In this way, if the Green's functions are restricted to horizontal layers, the translation invariance property holds separately for each of them. This means that multiple

Green's functions need to be computed to account for the different levels in which heat is dissipated and in which temperature is computed. More precisely, if the power is dissipated on N_p different layers and the temperature maps are required on N_t different layers, then $N_p \cdot N_t$ Green's functions need to be calculated. Each of them is a function of the horizontal spatial coordinates and depends on the temperature computation and power dissipation levels, i.e. it can be written as $G(x, y, z_j, t; x_0, y_0, z_i, t_0)$ where z_i is the level where power is dissipated and z_j the one in which temperature is computed. More precisely, since $G(x, y, z_j, t; x_0, y_0, z_i, t_0)$ are translation invariant along horizontal planes, the δ power can be assumed to be dissipated in $(x, y, t) = (0, 0, 0)$ and the notation $G_{z_i}(x, y, z_j, t)$ can be used to indicate the solution of

$$\rho(z)c(z)\frac{\partial}{\partial t}G_{z_i}(x, y, z_j, t) - \boldsymbol{\nabla} \cdot [k(z)\boldsymbol{\nabla} G_{z_i}(x, y, z_j, t)] = \delta(x, y, z_i, t). \tag{2.17}$$

$\theta_{z_i}(x, y, z_j, t)$, which is the temperature increase on level z_j due to power dissipated on level z_i, is computed as

$$\theta_{z_i}(x, y, z_j, t) = \int G(x, y, z_j, t; x_0, y_0, z_i, t_0)q(x_0, y_0, z_i, t_0)dx_0dy_0dt_0. \tag{2.18}$$

($a(\boldsymbol{\xi}) = 0$ because $\Omega = \mathbb{R}^n$). The total temperature increase on level z_j is computed exploiting the superposition principle as

$$\theta(x, y, z_j, t) = \sum_{i=1}^{N_p} \theta_{z_i}(x, y, z_j, t). \tag{2.19}$$

Since, in this way, the dependency of the differential operator L on z is accounted by using multiple Green's functions, the vertical dimension of the model can be finite. The impact of the top and bottom BCs at different distances from these boundaries are, indeed, already included in the different Green's functions. Normally, convective BCs are applied on top and bottom of the stack to simulate the heat transfer between the solid (3D-IC) and the fluid around it (air), which has its own temperature T_{amb}. However, since the operator L must be translation invariant in the x and y direction, the BCs should be independent of the x and y value, i.e. they should have constant coefficients.

Multiple thermal problems need, therefore, to be solved, one for each possible value of $i = 1, \ldots, N_p$ and $j = 1, \ldots, N_t$:

$$\rho(z)c(z)\frac{\partial T_{z_i}(x, y, z_j, t)}{\partial t} - \boldsymbol{\nabla} \cdot [k(z)\boldsymbol{\nabla} T_{z_i}(x, y, z_j, t)] = q(x, y, z_i, t), \qquad (x, y, t) \in \Omega \times \mathbb{R}^+ \tag{2.20a}$$

$$k(z)\frac{\partial T_{z_i}(x, y, z, t)}{\partial z} = h_t[T_{z_i}(x, y, z, t) - T_{amb}], \qquad z = z_t \tag{2.20b}$$

$$k(z)\frac{\partial T_{z_i}(x, y, z, t)}{\partial z} = -h_b[T_{z_i}(x, y, z, t) - T_{amb}], \qquad z = z_b \tag{2.20c}$$

$$T_{z_i}(x, y, z_j, 0) = T_{amb} \qquad (x, y) \in \Omega. \tag{2.20d}$$

where z_t and z_b are, respectively, the coordinates of the top and bottom boundaries while h_t and h_b are the values of the heat transfer coefficients applied on the top and bottom boundaries. It is important to note that the h_t and h_b coefficients are not a property solely on the surface but they depend mostly on the characteristics and flow of the fluid in contact to the surface [56]. This is one of the causes for nonlinearity in the BCs [44]. In this thesis the heat transfer coefficients are computed for each geometry and cooling solution but they are assumed independent on the dissipated power and ambient temperature (Section 5.2 and Paragraph *"Boundary conditions"* in Section 5.3.4). Due to the presence of the T_{amb} term in the BCs, multiple solutions of this problem cannot be superposed and, therefore, the convolution algorithm cannot be correctly applied.

However, if the temperature increases on each specific layer are considered, i.e.

$$\theta_{z_i}(x, y, z_j, t) = T_{z_i}(x, y, z_j, t) - T_{amb}, \tag{2.21}$$

the PDEs governing the evolution of $\theta_{z_i}(x, y, z_j, t)$ can be written as

$$\rho(z)c(z)\frac{\partial \theta_{z_i}(x, y, z_j, t)}{\partial t} - \nabla \cdot [k(z)\nabla \theta_{z_i}(x, y, z_j, t)] = q(x, y, z_i, t), \qquad (x, y, t) \in \Omega \times \mathbb{R}^+ \tag{2.22a}$$

$$k(z)\frac{\partial \theta_{z_i}(x, y, z, t)}{\partial z} = h_t \theta_{z_i}(x, y, z, t), \qquad z = z_t \tag{2.22b}$$

$$k(z)\frac{\partial \theta_{z_i}(x, y, z, t)}{\partial z} = -h_b \theta_{z_i}(x, y, z, t), \qquad z = z_b \tag{2.22c}$$

$$\theta_{z_i}(x, y, z_j, 0) = 0 \qquad (x, y) \in \Omega. \tag{2.22d}$$

In this way, therefore, the superposition principle can be directly applied. This means that the $G_{z_i}(x, y, z_j, t)$ functions computed for these problems can be used to calculate, by convolution, the temperature increases $\theta_{z_i}(x, y, z_j, t)$. Assuming that T_{amb} remains constant both in time and space, the term $a(\boldsymbol{\xi})$ in equation (2.5) is equal to T_{amb} and the temperature on each active layer is, then, computed as

$$T(x, y, z_j, t) = \sum_{i=1}^{N_p} \theta_{z_i}(x, y, z_j, t) + T_{amb}. \tag{2.23}$$

2.2.3 Steady state and transient methodology

In steady state the impulsive power generating the $G_{z_i}(x, y, z_j)$ functions is continuous in time, meaning that the system heats up until equilibrium. For the transient regime, instead, the power should also be impulsive in time, meaning that the system heats up and then cools down. More precisely $G_{z_i}(x, y, z_j, t)$ is the solution of

$$LG_{z_i}(x, y, z_j, t) = \delta(x, y, z_i)\delta(t). \tag{2.24}$$

Let's now define the function φ_{z_i} as the solution of

$$L\varphi(x,y,z_j,t) = \delta(x,y,z_i)H(t) \tag{2.25}$$

where $H(t)$ is the Heaviside function, meaning that the power is continuously dissipated from time 0 on, as for steady state. Since $\delta(t) = \frac{\partial H(t)}{\partial t}$, from equations (2.16), (2.24) and (2.25), we obtain

$$\begin{aligned} LG_{z_i}(x,y,z_j,t) &\overset{(2.24)}{=} \delta(x,y,z_i)\delta(t) = \delta(x,y,z_i)\frac{\partial H(t)}{\partial t} \overset{(2.25)}{=} \frac{\partial}{\partial t}L\varphi(x,y,z_j,t) \\ &\overset{(2.16)}{=} \frac{\partial}{\partial t}\left\{\rho(z)c(z)\frac{\partial}{\partial t}\varphi_{z_i}(x,y,z_j,t) - \boldsymbol{\nabla}\cdot[k(z)\boldsymbol{\nabla}\varphi_{z_i}(x,y,z_j,t)]\right\} \\ &= \rho(z)c(z)\frac{\partial}{\partial t}\frac{\partial\varphi_{z_i}(x,y,z_j,t)}{\partial t} - \boldsymbol{\nabla}\cdot\left[k(z)\boldsymbol{\nabla}\frac{\partial\varphi_{z_i}(x,y,z_j,t)}{\partial t}\right] \\ &\overset{(2.16)}{=} L\frac{\partial\varphi_{z_i}}{\partial t} \Rightarrow G_{z_i}(x,y,z_j,t) = \frac{\partial\varphi_{z_i}(x,y,z_j,t)}{\partial t} \end{aligned} \tag{2.26}$$

This means that it is possible to first compute the $\varphi_{z_i}(x,y,z_j,t)$ temperature responses for impulsive power dissipation in space and continuous in time. The subsequent derivative with respect to time provides the Green's functions $G_{z_i}(x,y,z_j,t)$ for impulsive power dissipation.

2.3 General assumptions for the FTM

In order to model the thermal behavior of a 3D package, some simplifications of the real device need to be made. In particular, to build a FTM based on the Green's functions theory, the following assumptions on the modeled structure and BCs are necessary.

Temperature independent material properties In order to apply the superposition principle and the Green's function theory, the partial differential operator L has to be linear. This means that the material properties are assumed to be temperature independent. According to [90], this assumption is acceptable if the difference between the maximum and the minimum temperature experienced during the considered phenomenon remains below $\sim 50°C$ (cf. Section 6.1).

Infinite horizontal dimensions In order to fulfill the horizontal translation invariance property of L, the lateral boundaries should not affect the temperature responses of the device, wherever the power is dissipated. This means that the system should not have lateral boundaries and, therefore,

it has to be infinitely large. This hypothesis has a big impact on the final temperature profiles, especially if power is dissipated close to the boundaries of the real device and/or if the applied cooling solutions on top and bottom of the stack are poor.

Homogeneous material layers Another condition to be able to fulfill the horizontal translation invariance property of L is the independence of the material properties of the horizontal position. This means that the modeled structure has to be constituted of stacked layers of homogeneous materials. In case of small embedded structures, equivalent material properties can be used. The inaccuracy due to this assumption might, in some situations, be considered as a second order effect. However, accounting for the specific position of the embedded microstructures may be relevant if their placement has to be thermally optimized (cf. Section 9.4.2).

Top and bottom BCs with equivalent constant heat transfer coefficients As well as the material properties, to satisfy the translational invariance property of L, also the BCs applied on top and bottom of the stack should be independent of the horizontal position. Since convective BCs are typically applied to model the effect of the cooling solution, the convective coefficients h_t and h_b have to be constant. This assumption limits the possibility to account for the thermal impact of the package. The package is, indeed, normally larger than the die stack itself, and, as a consequence, it allows the heat to spread outside the die stack. This heat spreading could be included in the thermal model by considering space dependent convection coefficients on the top and bottom of the die stack. As a consequence, the impact of assuming constant heat transfer coefficients may be relevant and depends on the kind of package and on the applied BCs (cf. Section 5.1).

Planar power dissipation Since the active regions in the dies are much thinner than the dies themselves (couple of nm versus tens of μm), they are assumed to be planar.

2.4 Limitations

All these assumptions cause, of course, limitations to what can be modeled by adopting this FTM strategy. Figure 2.2 shows the transformation from the schematic of a realistic packaged 3D-ICs to a structure that can be modeled by means of convolution operations between power maps and Green's functions, which are restricted to horizontal planes. In the following of this thesis, the term *package configuration* refers to a structure as the one on the left hand side while *stack configuration* to something similar to what is illustrated on the right (in fact, a finite size as large as the *die stack* is considered starting from Chapter 3). The term *die stack* is, instead, used to indicate the cuboidal section of the 3D-IC constituted by

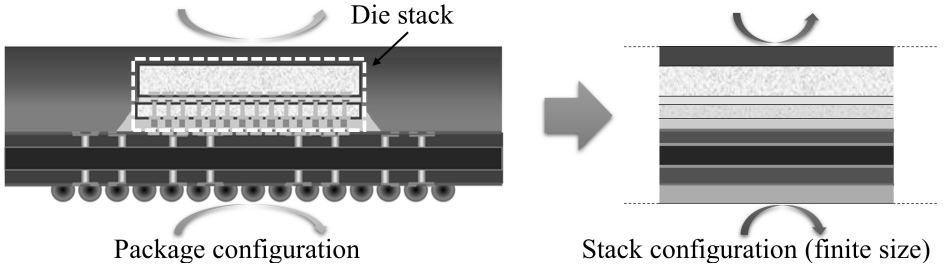


Figure 2.2: Sketch of a realistic packaged 3D-ICs (from [75]), on the left-hand side, and of the corresponding structure that can be modeled, according to the listed assumptions, by the Green's function approach, on the right-hand side.

the stack of dies, interface materials, BEOLs, ... This is the heart, the operating part, of the 3D-IC and it is normally enclosed in a package to protect it and to allow connections with the PCB and the outside world. The objective of the next chapters is to try to reduce the gap between the modeled structure and the more realistic schematic. The most significant constraints in thermal modeling of 3D-ICs via this FTM are listed hereafter.

Boundary conditions The constraints related to the BCs can be divided into two sets, depending whether they refer to lateral or top/bottom boundaries.

Lateral Since the vertical dimensions of the 3D-IC are normally much smaller than the horizontal ones (μm vs cm) and since the cooling solutions (heat sink, heat spreader, PCB, . . .) are applied on top and bottom of the stack, insulation is normally assumed on the lateral sides of the stack and of the package in classical thermal simulations [53]. In this FTM, the *method of images* is employed to include the adiabatic effect of the lateral boundaries and to allow the modeling of a finite dimensional structure. This methodology is illustrated in Chapter 3.

Top and bottom Equivalent convective BCs with appropriate constant heat transfer coefficients are applied on top and bottom of the stack configuration. The values of these coefficients are chosen to mimic the thermal resistance of the top part (normally mold compound, lid, heat spreader and convection to ambient) and of the bottom part of the heat path (normally substrate, solder, PCB and convection to ambient). As already stated in Section 2.2.2, their thermal impact is included during the computations of the temperature responses to HS power dissipation.

Thermal spreading This is a phenomenon happening when the heat is conducted from a smaller to a larger area (cf. Section 1.3.2). In a 3D-IC device, it mainly happens at two levels: in the die stack itself, i.e. from the heat sources to the stack of dies, and in the package, i.e. from the stack to the substrate, overmold or heat sink.

Stack Let us consider the stack configuration in which different layers can be of different materials. The heat spreading *within* the die stack is due to non uniform power dissipation on the active layers. If, in this set up, uniform power is dissipated, no thermal spreading happens. However, in case a more general non-uniform power map is proposed, the convolution algorithm is able to take into account the spreading that happens *within* the die stack due to this non uniformity. This information is, indeed, stored in the temperature responses to HS power dissipation.

Package Since the stack configuration does not include the full package, its thermal impact is not included in the model. Moreover, the package, the heat sink, the heat spreader, the PCB, . . . have a much larger footprint area than the stack itself, meaning that lateral heat spreading occurs when the heat passes through them, during its path from the dissipation level in the die stack, to the ambient. This issue is tackled in Chapter 5.

Heterogeneous materials Even if just the stack section of the device is considered (no package around) different materials may be present in one single layer. This can be the case, for example, of TSVs embedded in silicon dies or μbumps surrounded by undefill material.

Uniformly distributed If the small structures are approximately uniformly distributed, then homogenization techniques can be employed to compute equivalent material properties for each specific heterogeneous layer (cf. Paragraph *"Equivalent material properties"* in Section 5.3.4 for more details). A representative volume is considered and the equivalent thermal properties are extracted. The equivalent thermal conductivities, which can be orthotropic, are computed from the values of the thermal resistances obtained when the heat flow is constrained to one direction. Since the heat capacity C is a volumetric quantity, its equivalent value for each level of interest is computed by means of volume average. The corresponding values of the specific heat, c, and of the mass density, ρ, are extracted from it. This is possible since, in thermal simulations, just the product $c\rho$ plays a role, the two terms never appear separately. Making use of equivalent material properties, the global effect of these heterogeneities can be directly included in the FTM.

Specific layout The convolution based FTM cannot, however, directly include the local and the global thermal impact of microstructures. If the user wants, for instance, to compare various layouts in which the μbumps are not uniformly distributed, the FTM with homogenized material properties cannot give any relevant thermal indication (cf. Chapter 4).

Temperature dependency of material properties In general, material properties can be temperature dependent. In case of microelectronics packages, this is

of particular concern for silicon. First of all, power is dissipated in the silicon dies and, as a consequence, higher temperature gradients are experienced in this material. Second, the thermal conductivity of silicon strongly depends, with a negative feedback loop, on temperature. More precisely [82],

$$k_{Si}(T) = 148\left(\frac{300}{T}\right)^{1.65} W/mK \tag{2.27}$$

where T is in degree Kelvin. This means a reduction in thermal conductivity of almost 38% for a temperature variation from 28°C to 128°C. For the other materials in the 3D package this dependency is much lower (for copper, for example, it is around 1% for the same temperature variation [110]). However, the temperature dependency of the material properties cannot be directly included in this FTM since the whole theory, on which the model is based, has been developed for linear problems. As a consequence, constant values, based on the expected average operating temperature, are initially considered (cf. Chapter 6).

Temperature on horizontal layers The FTM provides the temperature maps on user defined horizontal levels. These are normally the same as the active layers as they provide the most relevant thermal information. Even if considering a large enough number of layers allows to have a 3D representation of the temperature distribution in the device, this FTM has not been designed to this aim.

2.5 Numerical implementation

The two commercial softwares Matlab [66] and Msc Marc [69] have been used for the numerical implementation of the FTM. The main part of the algorithm has been developed using Matlab while Marc has been used to solve all the FEM models that will appear in the following discussions. It has, in particular, been used to obtain the *reference FEM* results, against which the FTM has been validated.

In order to numerically compute the temperature profiles exploiting the Green's function theory, equation (2.18) has to be discretized. This means that the two fundamental ingredients, G and q, have to be discretized over appropriate grids (cf. Section 3.5) and that the integrals have to be transformed into sums. In the following, the nomenclature *hotspot response* (HSR) and *power map* (PM) are used to indicate, respectively, the discretized version of G and q. This is done in order to distinguish between the analytical, continuous quantities and the discrete ones.

2.5.1 Hot spot responses

The G functions were defined as the temperature responses of the system to δ heat sources. Since the HSRs are defined in the discretized domain, the δ heat sources are transformed in *unit cell* impulses, to mimic the infinitely small area of the δ functions, and the temperature responses are normalized with respect to the total dissipated power (or energy in case of transient regime), to mimic the unitary integral.

One of the main properties that the HSRs need to fulfill is the translation invariance on each horizontal level. This means that, if (x_{HS}, y_{HS}) indicates the location where the HS power is dissipated starting from time $t_0 = 0$, then $HSR(x, y, z_j, t; x_{HS}, y_{HS}, z_i, t_0) = HSR_{z_i}(d, z_j, t)$ where $d = \sqrt{(x - x_{HS})^2 + (y - y_{HS})^2}$ is the distance, on a fixed horizontal plane, between the location (x, y) where the HSR is computed and the center of the HS. For this reason, to compute the HSRs, 2D-axisymmetric models can be built and quickly solved by FEM. For each 3D-IC, the geometry of the model used to generate the corresponding HSRs is constituted by a stack of different homogeneous layers whose thicknesses and homogenized material properties are chosen to best represent the vertical cross section of the real 3D-IC stack. Concerning the horizontal direction, the modeled structure is much larger than the real stack. This is necessary to avoid the thermal impact of the insulating lateral boundary, ensuring, in this way, translation invariance of the HSRs. The top and bottom heat transfer coefficients are chosen equal to the ones that would be applied on top and bottom of the analogous (same vertical cross section and same, finite, horizontal size as the die stack) 3D-FEM model for the stack (cf. Paragraph *"Boundary conditions"* in Section 5.3.4). HS power is subsequently dissipated on each active layer and the temperature increases, which are 1D curves, are extracted on all the levels of interest. The HSRs for steady state are obtained by normalizing these curves with respect to the total dissipated power in the HS having, as a consequence, °C/W units. In case of transient regime, since also the time step is considered, the HSRs have units in °C/J (more explanation comes later in this Section).

Sources of inaccuracy in the HSRs

The finite element calculation of the HSRs introduces a number of errors that are due to:

Mesh the discretization of the domain in which the HSRs are computed and the intrinsic nature of FEM to approximate the solution on a finite dimensional space;

Lateral BC the lateral boundary effect. The model should be considered of infinite dimensions to ensure translation invariance of the HSRs. However, the

thermal impact of HS power dissipation decreases with distance. For this reason, the 2D-model can be of finite dimensions on condition that the lateral boundary is far enough not to have a significant impact on the temperature profiles;

δ **heat source** the G function should be computed for a δ power impulse. In the discrete domain, however, the δ function can not be represented. This means that the HSRs are basically unit-cell responses where power is dissipated over one cell in the discretized geometry. Accuracy issues may arise in the selection of the unit-cell and in the combination of the HSRs, which have circular symmetry, with the power map, which is defined on a square grid.

In the following, each of these possible sources of inaccuracy is further discussed.

Mesh

The first step needed to approximate the solution of a PDE by FEM is the discretization of the domain Ω in which the problem is defined. In case of HSRs, this is the 2D vertical cross section of the die stack. A mesh based on rectangular elements is considered. After the geometry has been created and meshed, a set of basis functions has to be chosen to approximate the solution of the PDE. The accuracy of the model depends on the combination of mesh (mainly) and basis functions. For this simple 2D-axisymmetric model, biquadratic basis functions, which provide higher accuracy than bilinear ones, have been selected and, as a consequence, each element is equipped with eight nodes (four corners and four middle points).

Once the basis functions have been selected, the mesh can still be changed to improve accuracy. The finer the mesh, the more accurate the solution. However, the computational time increases with the number of nodes. A good trade-off between accuracy and computational efficiency should be selected. A possible solution is to consider a non-uniform mesh. Smaller elements are used in proximity of *critical* regions and larger elements where the solution has little variations and there are no discontinuities. This is what has been done in this case: smaller elements are considered close to the areas where the material properties change and where the power is dissipated. These are the regions where higher variation in temperature is expected. Figure 2.3 shows a section of the mesh used to discretize the geometry of the model used to extract the HSRs referring to the schematic in the same Figure. In this example, a two dies stack is considered, different colors refer to different materials and the arrows indicate the positions where the HS power is applied. The number of layers, the thicknesses of the layers, the material properties, the BCs, the dimension of the HS and the location where the power is applied depend on the structure that is modeled.

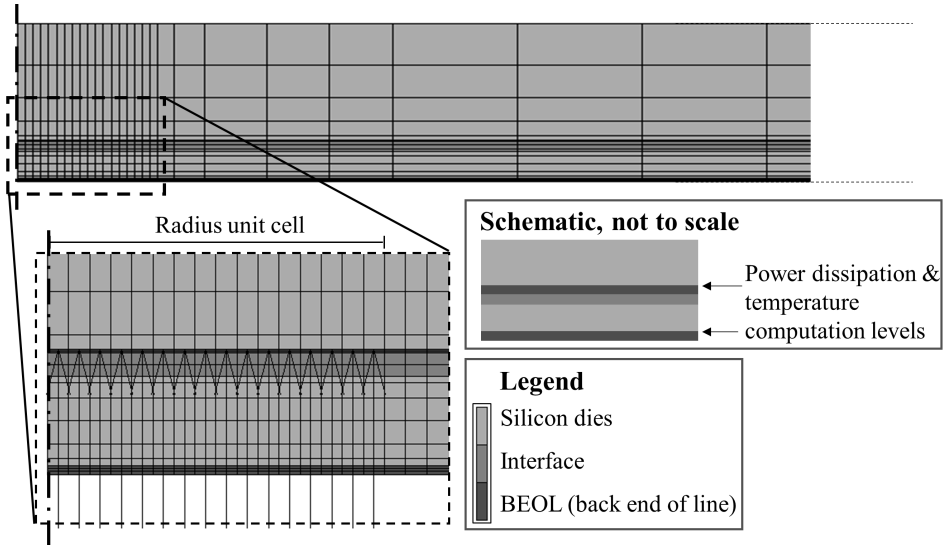


Figure 2.3: Meshed geometry of the 2D-axisymmetric model used to compute the HSRs. Different colors indicate different materials.

Moreover, since in general the accuracy of a FEM solution improves by reducing the element size, the *mesh independence* of the result has to be proven. This means to show that the effect on the solution of reducing the element size is negligible. The mesh independence for the model of the HSRs is shown in Figure 2.4, where the temperature has been computed using the mesh shown in Figure 2.3 (mesh $\bar{h}$) and halving it in both directions (mesh $\bar{h}/2$). The percentage error is plotted with diamond markers and refers to the right axis of the plot. Since the error is always lower than 0.025%, we can conclude that mesh independence is achieved. Moreover, the "higher" values refer to positions where the temperature is low and, therefore, they are mainly due to numerical rounding. Ad hoc methodologies have been developed to estimate the discretization error resulting from a specific mesh. However, they won't be discussed here since it is clear from the low value of the percentage error that the model is mesh independent [32,119]. Although the results in Figure 2.4 refer to a specific case, the validity of the mesh independence property for the model of the HSRs has been confirmed also for structures with different thicknesses of the layers, different material properties and boundary conditions. All the models for the HSRs, meshed according to Figure 2.3, are, indeed, quite similar to each other even if they correspond to different devices.

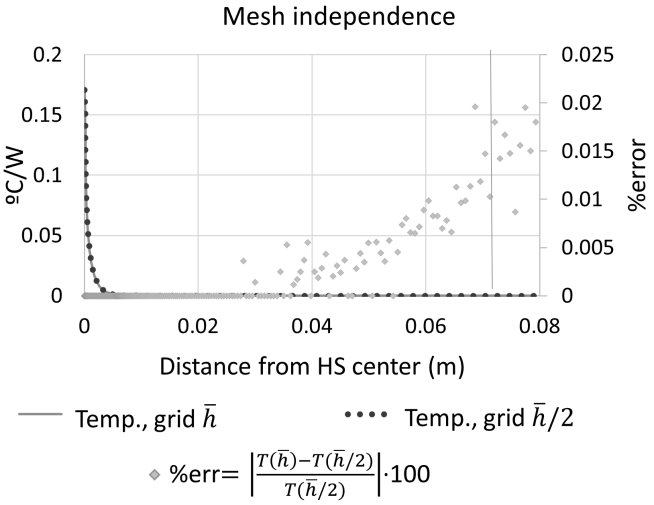


Figure 2.4: Mesh independence of the temperature profiles extracted from the HSRs (left axis). The profile depicted with the full line is obtained using the mesh illustrated in Figure 2.3 while the one indicated by a dotted line, by halving it. The percentage error is plotted with diamond markers and refers to the right axis.

Lateral boundary

The second source of error concerns the impact of the lateral BC on the values of the HSRs. The used FEM model has, indeed, a finite horizontal dimension and, as a consequence, the size of the geometry in the 2D-model should be such that, increasing it further, doesn't produce any difference in the HSR profiles. Otherwise, while convolving the HSR and the PM, the effect of this BC would be included in locations where it shouldn't and it would have a wrong impact on the overall temperature profiles. The distance from the HS center at which the impact of the lateral BC becomes negligible depends on the material properties, the geometry, the heat transfer coefficients used to model the BCs on the top and bottom of the stack and on the HS size. In Figure 2.5 the maximum value of the HSR is plotted on the left vertical axis as a function of the number of times the model is larger than the HS itself. On the right vertical axis, the difference in the maximum value of the HSRs obtained for different sizes of the model, $[HSR(x_{i+1}) - HSR(x_i)]/(x_{i+1} - x_i)$, is plotted. For the 2D-geometry modeled in this case, for example, a 2D-model that is thousand times larger than the HS ensures the HSRs not to be significantly influenced by the lateral BC. The data refer to steady state since this is the worst case scenario, the situation in which the heat reaches its maximum spreading. In the following, since the 2D model for the HSRs runs fast and since the distance from the HS center at which the lateral BC doesn't influence the value of the HS is case dependent, a model that is 2000 times larger than the HS has normally been

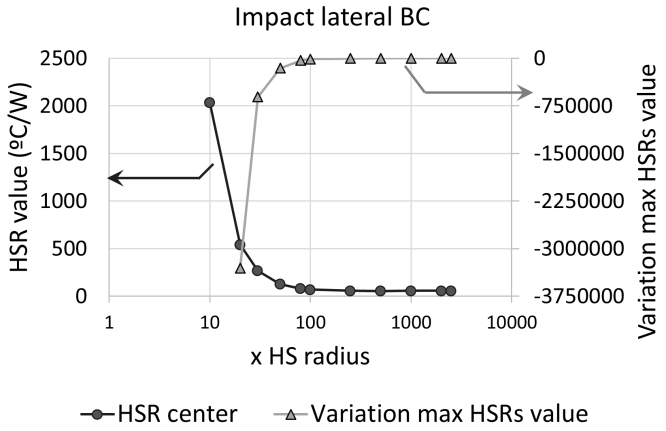


Figure 2.5: Impact of the lateral BCs on the peak value of the HSR. The max value of the HSR is plotted on the left vertical axis as a function of the number of times the model is larger than the HS itself. On the right axis, the difference in the max value of the HSR obtained for different sizes of the model, $[HSR(x_{i+1}) - HSR(x_i)]/(x_{i+1} - x_i)$, is plotted.

considered.

δ power source

The third source of error concerns the δ power source that generates the HSRs. In the discrete space, the δ function is approximated by a unit-cell heat source. Its size, $\bar{h}_{\text{HS}}$, should be the same as the grid size, $\bar{h}$, used for the power map. The three cases in which $\bar{h}_{\text{HS}} = \bar{h}$, $\bar{h}_{\text{HS}} > \bar{h}$ and $\bar{h}_{\text{HS}} < \bar{h}$ are tackled hereafter and illustrated in Figure 2.6 for a 1D geometry. In each of the three blocks, representing the three different cases, three graphs are shown (not to scale). The ones on the top left represent the HSR and the yellow area indicates the area where the unit cell power, which generates the HSR, is applied. In the graphs concerning the PMs, the purple lines indicate the values of the dissipated power while the green areas represent the region where power is applied. The grid sizes are also reported. The graphs on the second line of each block report the temperature increase, obtained by superposing (or convolving) the HSRs and the PMs. The regions corresponding to the locations where the power in the PM (green) and where the generating power in the HSRs (yellow) are dissipated, are also indicated.

Applying superposition, copies of the HSRs are superposed according to the dissipated power information stored in the PM. In case $\bar{h}_{\text{HS}} = \bar{h}$ (top-left corner of Figure 2.6), the region where the power generating the HSR is dissipated coincides with the region where the power is dissipated in the PM. In case $\bar{h}_{\text{HS}} > \bar{h}$ or

$\bar{h}_{\mathrm{HS}} < \bar{h}$, this is not the case. In the top-right corner of Figure 2.6, the case in which $\bar{h} = 1/3\bar{h}_{\mathrm{HS}}$ is shown. In such a situation there are some areas in which the temperature is computed as if the power were dissipated three times and there are grid elements, in the area in which the power is not dissipated according to the PM, where, in fact, while applying superposition, it happens to be. The last possibility, with $\bar{h} = 3\bar{h}_{\mathrm{HS}}$ is illustrated in the bottom section of Figure 2.6. In this case, while applying superposition between the HSR and the PM, it is as if the power is dissipated just in the central part of the effective dissipation area. Therefore, $\bar{h}_{\mathrm{HS}} = \bar{h}$ should be chosen.

2D-HSRs

In the previous discussion about $\bar{h}_{\mathrm{HS}}$ and $\bar{h}$, both the PMs and the HSRs were one-dimensional. When the three dimensional domain is considered and the active layers are two dimensional, the correspondence between the size of the HS, $\bar{h}_{\mathrm{HS}}$, and the grid size, $\bar{h}$, is not so straightforward. The PMs are, indeed, discretized by rectangular cells in the *Cartesian* coordinate system while the HSRs, due to the axisymmetrical nature of the FEM model, are originally discretized by rectangular cells in the *polar* coordinate system. In order to convolve the HSR with the PM, both quantities need to be considered in the same coordinate system. Taking into account the real shape of the 3D packages, the Cartesian system is more appropriate. This means, however, that $\bar{h}_{\mathrm{HS}}$ refers to a *diameter* while $\bar{h}$ to the edge of a square. Since there is no way to cover a square area using a circle without having superposition or uncovered space and since the power is dissipated over *areas*, we consider the situation in which the area where the HS power is dissipated is as large as the area represented by one grid elements in the PM. This means that $r_{\mathrm{HS}} = \frac{\bar{h}_{\mathrm{HS}}}{2} = \frac{\bar{h}}{\sqrt{\pi}}$, where r_{HS} is the radius of the HS in the 2D-axisymmetric model. Another option [37] would be to use a 3D-model to generate the HSRs, dissipating power in a unit square cell of size $\bar{h}$. In this way, however, the computational time for the FEM increases significantly without a significant improve in accuracy.

The HSRs obtained by the axisymmetric FEM models are 1D curves but they have to be convolved with matrices storing the information about the dissipated power over the active areas, which are 2D surfaces (cf. Section 2.5.2). This means that a 2D spatial representation of each HSR is needed. Square matrices are created and, to each of their cells, the value of the HSRs corresponding to the distance between that specific cell and the center of the matrix is assigned. It is important to note that the HSRs need to have an odd number of rows and columns in order to be able to store the peak temperatures in their centers. Linear interpolation is used to get the values in locations other than the nodes in the FEM. The process to generate the 2D matrices that spatially represent the HSRs is illustrated in Figure 2.7.

It is important to stress that the HSRs depend on the geometry, the material

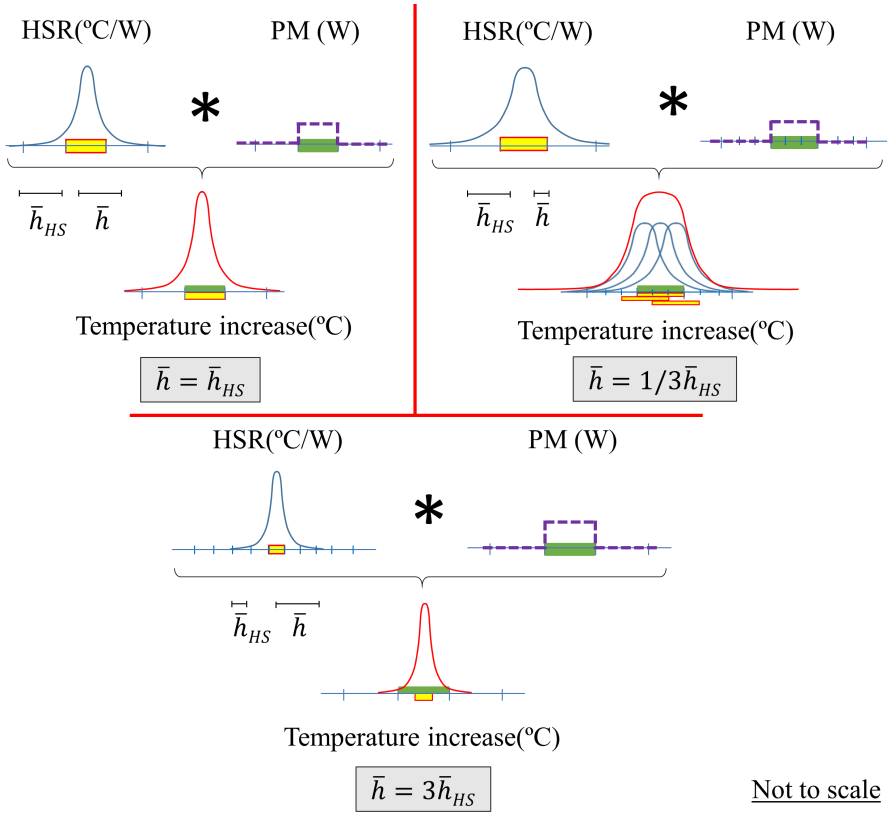


Figure 2.6: Importance of selecting the dimension of the HS, $\bar{h}_{\mathrm{HS}}$, equal to the grid size, $\bar{h}$, chosen for the PM. The cases for $\bar{h} = \bar{h}_{\mathrm{HS}}$, $\bar{h} = 1/3\bar{h}_{\mathrm{HS}}$ and $\bar{h} = 3\bar{h}_{\mathrm{HS}}$ are illustrated, respectively, on the top-left, top-right and bottom part of the Figure.

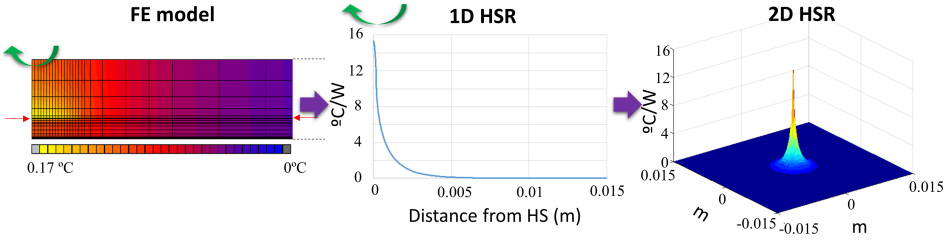


Figure 2.7: Process to generate the 2D matrices representing the spatial variation of the HSRs starting from the 2D-axisymmetric FEM temperature results.

properties and the cooling solution applied on top and bottom of the stack. This means that if one of these parameters changes, the HSRs need to be recomputed. Luckily, the related FEM models run really fast, since they are two dimensional. This is one of the reasons why 2D-axisymmetric models, with circular power sources, have been preferred to 3D-models, with square power sources, to generate the HSRs. Moreover, coupling the results obtained by the 2D-FEM axisymmetric models with the planar discretization of the PMs, allows combining the effect of both a fine vertical mesh (in the FEM) and a fine horizontal mesh (in the PMs), without an actual discretization of the full 3D geometry. This ensures high accuracy of the temperature profiles.

Transient HSRs

For steady state simulations, $N_p \cdot N_t$ 2D-HSRs need to be extracted, where N_p is the number of active layers and N_t the number of layers in which the temperature profiles have to be computed. For transient simulations, also the time variable has to be considered. What happens is that the 1D transient temperature profiles, obtained by the 2D-axisymmetric transient FEM model, are extracted at each simulated time step. The sizes of the time steps, which are non-uniform, are automatically selected by the FEM software according to the temporal variations of temperature and the simulations are run until steady state is reached (cf. Section 3.6). However, as it will be explained later (cf. Section 2.5.3), the convolution algorithm works with a constant time step. This means that the temperature profiles obtained by FEM are extracted and spline interpolation in time is employed at each spatial position to obtain the HSRs at the needed points in time. To each of these 1D temperature profiles, the same procedure used in steady state to obtain 2D-HSRs is individually applied.

All these obtained matrices, referring to the same temperature computation and power dissipation levels, can be grouped in a 3D matrix in which the rows and columns represent the spatial variation at a particular point in time and the third dimension represents the evolution in time of a particular location in space. This means that in transient regime $N_p \cdot N_t$ 3D-HSRs are required. Moreover, since, in FEM, it is computationally easier and faster to obtain the temperature profiles for power dissipation continuous in time, the identity in equation (2.26) is exploited. This means that the temperature results, extracted from the FEM model and obtained for continuous power dissipation, are both numerically differentiated with respect to time, to get the HSRs for impulsive power dissipation, and transformed into 3D matrices. This results in quantities with units $°C/J$.

2.5.2 Power maps

The power maps are matrices storing the information concerning the dissipated power on each active layer. A uniform discretization has to be selected for the PMs and the total power dissipated in the real device, in the area corresponding to a particular cell, is assigned to that cell. In case of steady state simulations, N_p 2D-PM matrices, with units of W, are defined, one for each active layer.

For transient simulations, 3D matrices are used in which, similarly to what happens for the HSRs, the time evolution of the dissipated power is stored in the third dimension. Moreover, assuming a constant time step Δt, the power map concerning a fixed point in time is assumed to be dissipated for a constant time Δt. For this reason, the dissipated power is multiplied by Δt and the transient PMs have units in Joule.

2.5.3 Convolution

As previously stated, the temperature profile on level z_j due to power dissipated on level z_i can be computed as

$$\begin{aligned} T_{z_i}(x,y,z_j,t) &= \int G_{z_i}(x-x_0, y-y_0, z_j, t-t_0) q(x_0, y_0, z_i, t_0) dx_0 dy_0 dt_0 + T_{amb} \\ &= G * q + T_{amb} = q * G + T_{amb} \\ &= \int G_{z_i}(x_0, y_0, z_j, t_0) q(x-x_0, y-y_0, z_i, t-t_0) dx_0 dy_0 dt_0 + T_{amb} \end{aligned} \tag{2.28}$$

since the convolution operator is commutative. To solve the problem numerically, also the convolution operator has to be discretized, not only the dissipated power and the G functions. The resulting temperature profiles will be matrices with the same resolution as the PMs and the HSRs. The steady state and the transient regimes are treated separately.

Steady state

In steady state, the temperature in cell $(\bar{i}, \bar{j})$ on level z_j can be computed by discretizing the two dimensional spatial convolution as

$$T_{z_i}(\bar{i}, \bar{j}, z_j) \approx \sum_{m=-a}^{a} \sum_{n=-b}^{b} HSR_{z_i}(m, n, z_j) PM_{z_i}(\bar{i}-m, \bar{j}-n) + T_{amb} \tag{2.29}$$

where $(2a+1)$x$(2b+1)$ are the dimensions of the HSR, whose peak corresponds to cell $(0,0)$, and, if for some n or m values the PM is undefined, it is taken equal

to zero. From this equation, the necessity to have a regular spatial discretization, which is the same for the HSR and the PM, becomes clearer. Because of the discussion on the δ power source in Section 2.5.1, it is, indeed, necessary that, every time $HSR_{z_i}(m,n,z_j)$ is multiplied by $PM_{z_i}(\bar{i}-m,\bar{j}-n)$, these terms refer to the same area. Since, all possible combinations between cells in the HSR and in the PM are present in equation (2.29), a fixed and equal spatial grid size is needed for the PM and the HSR.

According to equation (2.29), to obtain the temperature in a single point a double sum has to be handled. This results in a computational effort of $O(N^2)$, where $N = N_r N_c$ is the number of elements in the considered PM while N_r and N_c are, respectively, the numbers of rows and columns in the PM. When, in Chapter 3, the method of images is introduced to account for the finite dimension of the die stack, N, N_r and N_c refer to the extended PM, including images. For fine resolutions or large geometries, the time needed to obtain the result can be high. Luckily, a discretized version of the convolution theorem stated in Section 2.2 holds.

The *discrete Fourier transform* (DFT) is an invertible, linear transformation $\mathcal{F}_D : \mathbb{C}^N \to \mathbb{C}^N$ with the property of *completeness*. This means that no particular properties have to be fulfilled by the functions themselves to allow direct and inverse transformation. For any $N > 0$, an N-dimensional complex vector has both a DFT and an inverse DFT (IDFT).

Theorem 2. *The* convolution theorem for the discrete Fourier transform *in* $\mathbb{C}$ *states that the convolution between two finite sequences* $\{\boldsymbol{x}\}$ *and* $\{\boldsymbol{y}\}$ *of length N can be obtained as the IDFT of the product of their individual DFT:*

$$\mathcal{F}_D^{-1}\{\boldsymbol{X}\cdot\boldsymbol{Y}\}_n = \sum_{l=0}^{N-1} x_l (y_N)_{n-l} \doteq (\boldsymbol{x} * \boldsymbol{y}_N)_n$$

where $\boldsymbol{X} = \mathcal{F}_D\{\boldsymbol{x}\}$, $\boldsymbol{Y} = \mathcal{F}_D\{\boldsymbol{y}\}$, x_l *is the* l-th *element in the sequence* $\{\boldsymbol{x}\}$ *and* $(\boldsymbol{y}_N)_n = y_{n(modN)}$, *with* $\{\boldsymbol{y}_N\}$ *the extension of* $\{\boldsymbol{y}\}$ *by periodic summation.*

This happens for Nx1 sequences. The multidimensional DFT can be computed by the composition of a sequence of 1D-DFT along each dimension. It has, therefore, the same properties as the 1D-DFT and an analogous multidimensional convolution theorem for the discrete multidimensional Fourier transform holds. The implementation of the DFT usually employs FFT algorithms that allow high reduction of computational time. The calculation of $\mathcal{F}\{\boldsymbol{x}\}$, strictly following the definition of DFT, requires, indeed, an effort proportional to $O(N^2)$. Several FFT algorithms have been proposed and they are able to reduce the computational time up to $O(N \log N)$ [11, 106]. This means a speed up in the range of $N/\log N$. The effective speed-up, however, depends on the prime factorization of N. In the most commonly used Cooley–Tukey algorithm, the best speed up is obtained if N is a power of two.

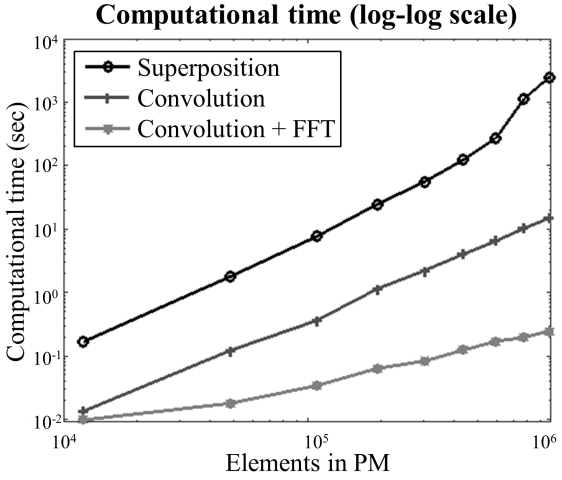


Figure 2.8: Computational time needed to apply the superposition principle by using superposition, convolution and convolution plus FFT as computational methods for different dimensions of the PM. Note that the graph is in log-log scale.

The computational time needed to apply the superposition principle by using 1) superposition, 2) convolution and 3) convolution plus FFT as computational method is shown in the log-log plot in Figure 2.8 for different number of elements in the PM and HSR matrices. The speed up achieved by convolution plus FFT is clearly visible.

Transient

For transient regime, two different calculation methodologies can be considered. The first one is based on convolution in space and superposition in time. Equation (2.28) can, indeed, be considered separating the space variables, $\boldsymbol{x} = (x, y)$, and the time variable t:

$$T_{z_i}(x, y, z_j, t_f) = \int_0^{t_f} \left[\int_\Omega G_{z_i}(x_0, y_0, z_j, t_0) q(x - x_0, y - y_0, z_i, t - t_0) dx_0 dy_0 \right] dt_0 + T_{amb}.$$

For a fixed value of t_0, the space integral is a convolution integral between the G function at time t_0 and q at time $t - t_0$. It can be seen as a *steady state* temperature increase profile at time t and on level z_j, $\bar{\theta}_{z_i}(x, y, z_j, t; t_0)$, due to impulsive power dissipated on level z_i at time $t - t_0$. Since a t_0 delay from the power dissipation moment is considered, the G function at time t_0, which represents the response of the system after that time interval, is used in the convolution. The internal convolution integral can be discretized similarly to the steady state regime, while

the external time integral using, for example, the rectangle method. This results in

$$\bar{\Theta}_{z_i}(\bar{i},\bar{j},z_j,\bar{t}_k;\bar{t}_l) \approx \sum_{m=-a}^{a}\sum_{n=-b}^{b} HSR_{z_i}(\bar{i}-m,\bar{j}-n,z_j,\bar{t}_l)PM_{z_i}(m,n,\bar{t}_k-\bar{t}_l) \tag{2.30}$$

and

$$T_{z_i}(\bar{i},\bar{j},z_j,\bar{t}_k) \approx \sum_{\bar{t}_l=1}^{\bar{t}_k} \bar{\Theta}_{z_i}(\bar{i},\bar{j},z_j,\bar{t}_k;\bar{t}_l) + T_{amb} \tag{2.31}$$

where $\bar{t}_k \in \mathbb{N}$ indicates the time step ($t_k = \bar{t}_k \Delta t$ is a point in time (*sec*)). This basically means that the transient temperature response can be computed as the superposition of the thermal impacts due to power dissipated in the past. These thermal impacts are computed taking into account how far in the past the power has been dissipated with respect to the present time t_k.

The second calculation methodology for transient thermal modeling is based on 3D-convolution in two spatial and one temporal variables. This can be done considering the time variable in the same way as the spatial ones and discretizing the convolution operator as for the steady state case. There is, however, a fundamental difference between the space variables and the time variable. Supposing to dissipate a unit cell, impulsive power in the system's origin at $t = 0$, then $HSR_{z_i}(x,\cdot,z_j,\cdot) = HSR_{z_i}(-x,\cdot,z_j,\cdot)$, $HSR_{z_i}(\cdot,y,z_j,\cdot) = HSR_{z_i}(\cdot,-y,z_j,\cdot)$ but $HSR_{z_i}(\cdot,\cdot,z_j,t) \neq HSR_{z_i}(\cdot,\cdot,z_j,-t)$. The value of $HSR_{z_i}(x,y,z,t)$ is actually defined only for $t \geq 0$, i.e. after power dissipation. However, due to the implementation of the convolution algorithm, the integration intervals are symmetric with respect to $t = 0$, meaning that both $HSR(\cdot,\cdot,t)$ and $HSR(\cdot,\cdot,-t)$ are needed. Before power is dissipated (future time, negative values of the time variable) the temperature response is zero everywhere. If $\bar{t}_{ss}$ represents the number of time steps considered in the HSRs, this means that $\bar{t}_{ss} - 1$ zeros matrices have to be added to the HSRs sets to cover the future times and to place the dissipation time in the center of the HSR time line. In this way, the location, both in space and time, of the power dissipation corresponds to the center of the 3D-HSRs and the 3D-convolution algorithm can be applied. Concerning $\bar{t}_{ss}$, as it will be shown in Section 3.6, this number is related to the time needed to the system to reach the steady state regime.

For both implementations, a fixed time step is required. The reason is analogous to the one for which a regular spatial grid is needed. To obtain the discrete temporal evolution of the temperature profiles, indeed, both $HSR_{z_i}(\cdot,\cdot,z_j,t_l)$ and $PM_{z_i}(\cdot,\cdot,t_k - t_l)$ are needed for each possible combination of t_k and t_l. Since, the HSRs and the PMs should exist for all these combinations in a consistent way, the time step has to be fixed. In other words, for each 2D-PM carrying the information about the power dissipated in the device a certain amount α of seconds before the actual time at which T is being computed, the corresponding HSRs referring to the response of the system exactly α seconds after HS power dissipation has to exist.

Figure 2.9 graphically illustrates the convolution procedure, with respect to the time variable, to obtain the temperature in one fixed location. The power dissipation

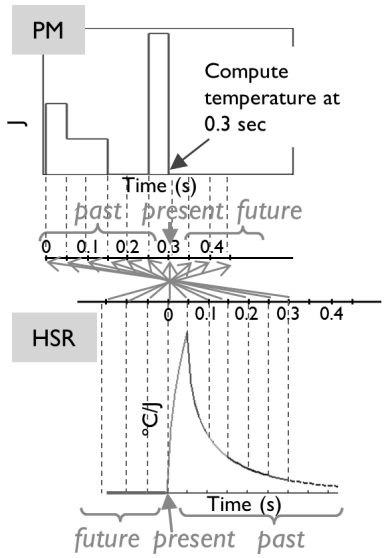


Figure 2.9: Illustration of the procedure for convolution in time: in this example just one point in space is considered.

profile is shown in the top plot while the HSR in the bottom one. Discretization with respect to the time step is also indicated. In this example, the temperature needs to be computed at $t_k = 0.3\ sec = \bar{t}_k \Delta t = 6 \cdot 0.05\ sec$, which is indicated as present time for the PM. This corresponds, therefore, to $t_l = 0\ sec$ in the HSRs graph. The past of the power dissipation (left side of the PM graph) needs, then, to be combined with the corresponding part of the HSR that indicates what happens *after* power dissipation (right side of the HSR graph). This is because, if a certain amount of power has been dissipated at $t_k - t_l$, with $t_l \neq 0$, then it has been dissipated a time t_l before present and, so, the corresponding HSRs value should be taken a time t_l after power dissipation. Since in the HSRs the dissipation time corresponds to 0, the value to be combined with power dissipated at $t_k - t_l$ in the PM corresponds to time t_l in the HSR graph. This procedure is repeated for all possible values of t_k in which the temperature has to be computed. As explained above, to allow for convolution in time, the value of the HSR corresponding to the moment when the HS power is dissipated has to be in the center of the HSR vector. As a consequence, $\bar{t}_{ss} - 1$ zero values need to be added to the left hand side of the HSR, to model future times.

This example doesn't consider space convolution. When the spatial information is included, 3D-convolution requires extra 2D zeros *matrices* instead of zero values. However, despite the increased dimension of the matrices to be convolved, the 3D-convolution algorithm turns out to be much faster than the one based on spatial convolution and time superposition, if the number of considered time steps $\bar{t}_f$

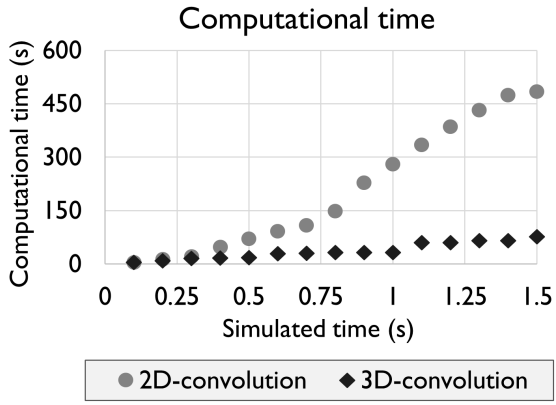


Figure 2.10: Comparison between the computational time needed for 3D-convolution and for 2D-convolution with subsequent time superposition.

is large enough. The exact number depends on the spatial resolution but it is relatively low (around 6 and 8). In most realistic situations, larger simulation intervals are considered for which, therefore, 3D-convolution is faster. The overall speed-up depends on the dimensions of the HSRs and PMs matrices.

Figure 2.10 shows a comparison between the computational time needed by the 3D-convolution algorithm and by the 2D-convolution plus subsequent time superposition one. The tic-toc Matlab command has been used to obtain the computational time. The results are obtained for a two die stack with a spatial discretization of 550x550 cells. A 50 msec time step is considered and the comparison has been performed for different simulated times ($2 \leq \bar{t}_f \leq 30$). The time length of the HSRs is always equal to the number of time layers in the PMs. As we can see from the graph, for small $\bar{t}_f$ values the performances of the two algorithms are comparable but, for larger simulated times, the 3D-convolution approach outperforms the 2D one. An improvement of a factor larger than five is reached, for example, considering 30 time layers, which correspond, in this case, to 1.5 sec of simulated chip activity. It is worth noting that the computational time doesn't grow linearly or quadratic with the number of time layers, $\bar{t}_f$. This is related to the convolution algorithm whose execution time depends on the prime factors of the dimensions of the involved matrices [65]. The algorithm performs, indeed, much better for powers of two or for decomposition with small prime factors. In the considered implementation, the matrices are padded in each individual dimension until a number of elements corresponding to a power of two is reached. Considering the computational time corresponding to the 3D-convolution, the presence of small "jumps" in the graph in correspondence of a total simulated time of 0.6 sec and 1.1 sec, for instance, are due to a change in the considered power of two.

It is important to note that the computational time of the FTM is not affected by the complexity of the PM but just by the number of elements in the convolved matrices. In FEM simulations, instead, the computational time depends on the complexity, both in time and space, of the PMs. In this kind of simulations the time step, in particular, can be automatically adapted according to the temporal variations of the power dissipation: to accurately model fast changes in the temperature responses, smaller time steps are selected. This means that the more the PMs vary, the higher the temperature gradients are, the smaller time steps are needed in FEM and the higher computational time is required. The time step in FEM models can also be fixed but this means that, to achieve at least the same accuracy as in the adaptive case, the smallest time step has to be used also when the temperature variation is slow. In case of this FTM, instead, the fast temperature changes have already been accounted for during the calculation of the HSRs. They are, therefore, already included in the HSRs and they don't need to be followed again for each specific PM. The size of the fixed time step in the FTM can be selected according to the length of the shortest interval in which the PM is kept constant. The improvement in computational time from FEM to FTM simulations depends, thus, on the PMs and it is larger the more complex the PMs are.

2.5.4 Flowcharts of the FTM algorithms

The flowcharts reporting all the steps of the algorithms developed for the steady state and for the transient FTMs in case of structures with infinite large horizontal dimensions are presented, respectively, in Figures 2.11 and 2.12. Boxes with rounded corners are used for inputs and outputs while rectangles with thick borders indicate blocks in which computations are performed. A gray background is used to indicate that the computations are performed by FEM (Msc Marc [69]), while a white background that they are performed by Matlab [66]. Similar flowcharts will be reported in the following of the thesis, both for steady state and transient regime, when particular features are introduced in the algorithm.

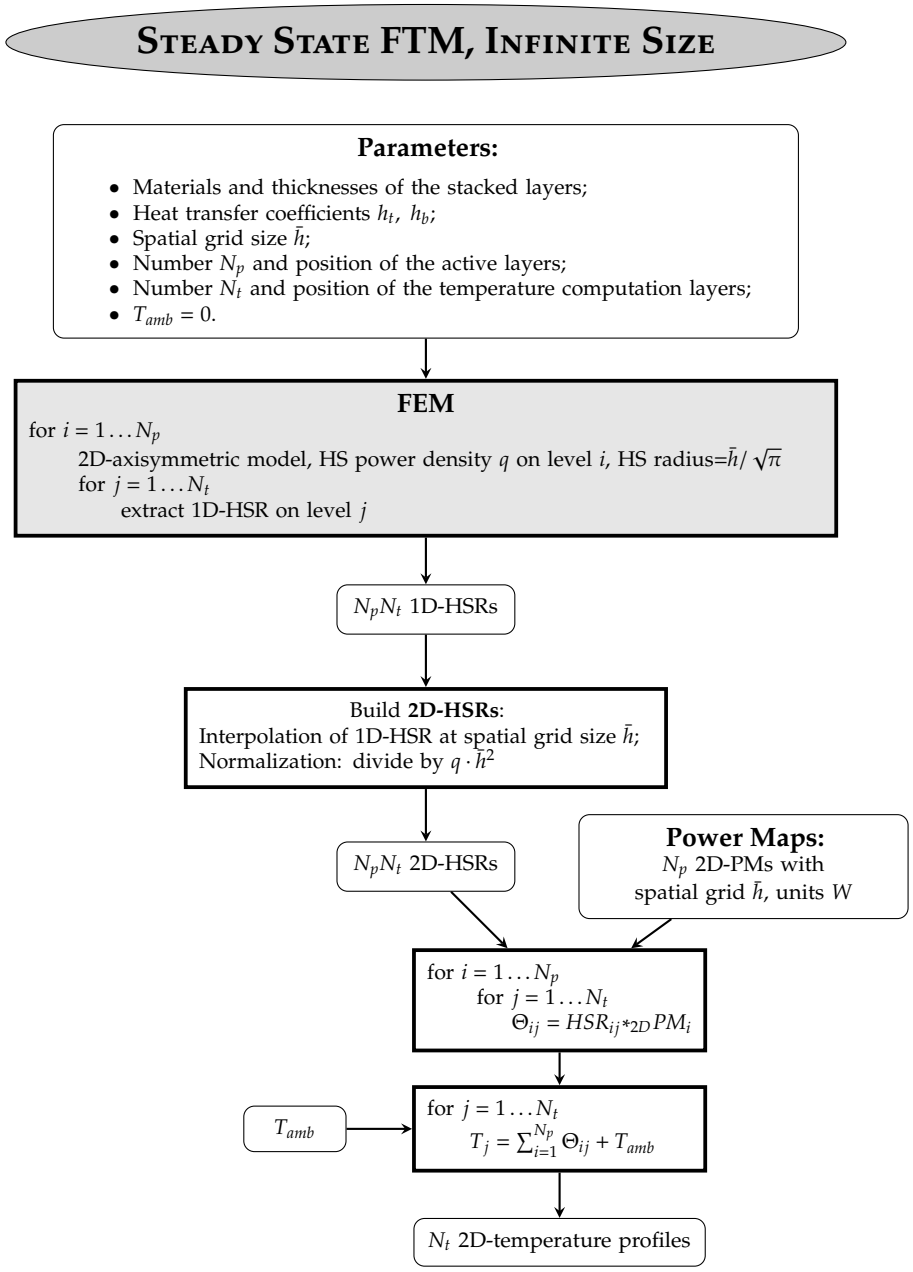


Figure 2.11: Flowchart representing the algorithm implemented for the steady state fast thermal modeling of 3D-stacks of infinitely large size.

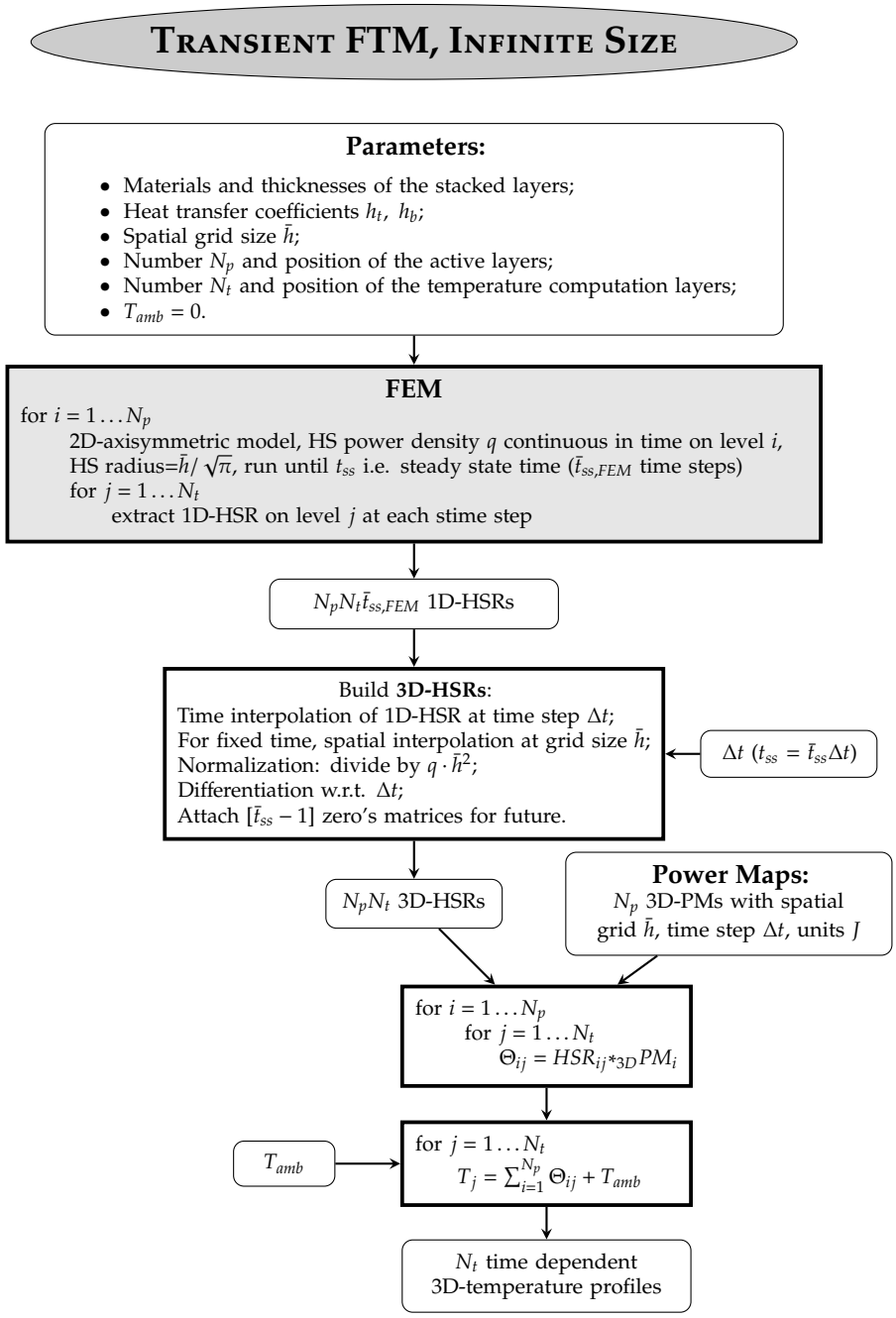


Figure 2.12: Flowchart representing the algorithm implemented for the transient fast thermal modeling of 3D-stacks of infinitely large size.

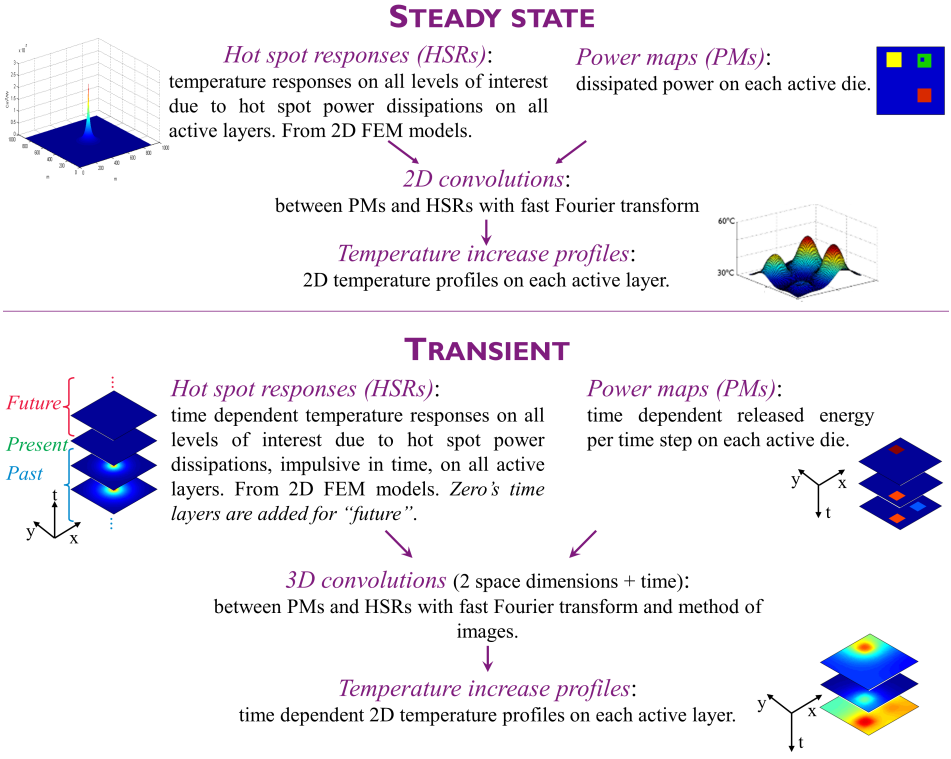


Figure 2.13: Illustration of the algorithms used to compute the steady state and transient temperature increase profiles using the convolution based FTM for infinite structures.

2.6 Summary

In this Chapter the methodology that forms the basis of the convolution based FTM for infinite structures has been presented. It allows the temperature computation, along horizontal levels, in case of stack configurations in which homogeneous horizontal layers are placed on top of each other. The model assumes temperature independent material properties as well as infinite horizontal dimensions of the modeled geometry, convective BCs with constant heat transfer coefficient on the top and bottom boundaries of the stack and planar power dissipation.

The properties and the procedure for the creation of the HSRs and of the PMs, which are needed to compute the temperature profiles, have been described in this Chapter. Figure 2.13 illustrates the ingredients and the steps needed to compute the steady state and the transient temperature increase profiles using the proposed FTM. For steady state, 2D matrices representing the HSRs and the PMs are needed while, for transient simulations these quantities are three dimensional because the

time variable has also to be considered. In the transient models, moreover, 2D zeros matrices have to be added to the HSRs to represent future times, so that the *present* is always in the middle of the third dimension. Doing so, a 3D-convolution approach can be implemented, which drastically reduces the computational time with respect to 2D-convolution.

Chapter 3

Modeling 3D Stacks of Finite Dimensions

3.1 Introduction

As previously stated, in order to achieve translation invariance of the HSRs, they should be computed for infinitely long structures. Geometries that are 2000 times larger than the HS are normally used to this scope and, as a consequence, the HSRs refer to structures much larger than the actual stacks. This means that the temperature results, obtained by convolving HSRs and PMs, refer to infinitely large geometries. In this Chapter, the *method of images* is presented as a solution to overcome this limitation and to allow the modeling of stacks of multiple homogeneous layers with *finite* horizontal size [24, 37]. Considering the real finite size of the stack rather than an infinite large structure has two main impacts on the temperature profiles. First, the *average* temperature increase is higher due to less room available for spreading and, second, the *local* thermal response depends on the horizontal position where power is dissipated. Power dissipation in the corner of the die results in higher temperature increase than if the same power source would be applied in the center of the die. The method of images takes into account both phenomena. At the end of this Chapter, therefore, the thermal analysis of a geometry as the one illustrated on top of Figure 3.1, with horizontal dimensions as large as the die stack itself, can be performed, allowing for a more realistic thermal characterization of the 3D-ICs. The main steps of this algorithm are illustrated in the flowchart in Figure 3.1. The step that is new with respect to the flowchart in Figure 2.1 and that allows considering the finite dimension of the modeled structure is highlighted in gray.

In the first part of this Chapter the mathematical derivation of the method of

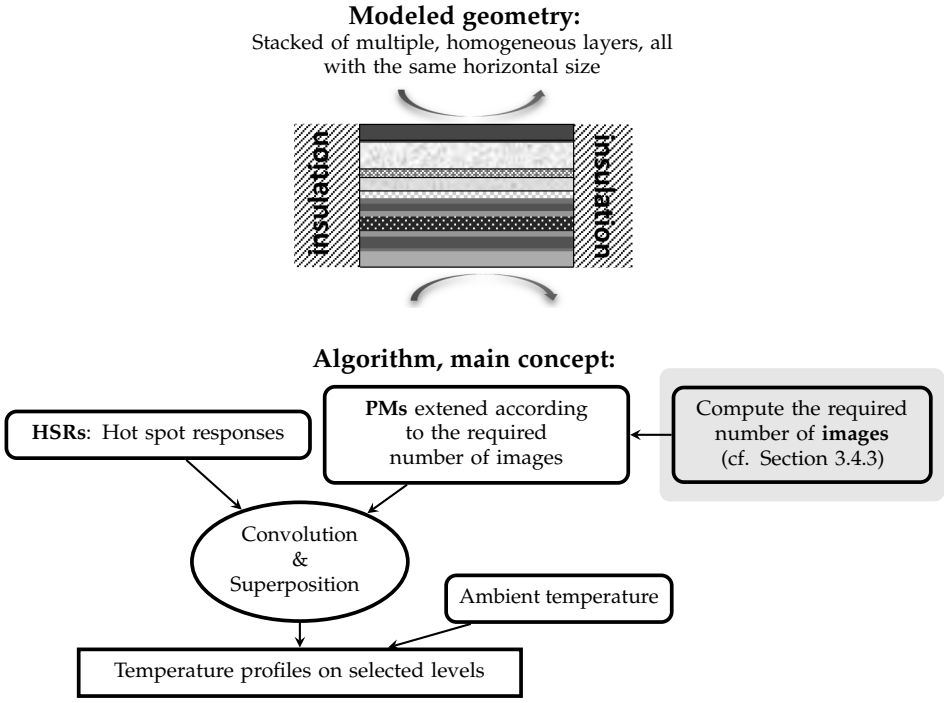


Figure 3.1: Modeled geometry and main concept of the algorithm described in this Chapter. The section specifically introduced and discussed in this Chapter is highlighted.

images is presented. Later, an accuracy assessment is performed on the number of images required to accurately model the effect of the boundaries. This is because, as it will be shown, from a theoretical point of view an infinite number of images is needed to convert the thermal response of an infinitely large structure into a finite one. In Section 3.4.1, a faster method, the *annulus method*, is proposed to compute the steady state temperature increase on layer z_j due to uniform power dissipation on level z_i, starting from the information stored in $HSR_{z_i}(x, y, z_j)$. This algorithm is used multiple times throughout the thesis. In this Chapter, it forms a fundamental part of the developed accuracy assessment. Other two accuracy-related topics, which are discussed hereafter, are the selection of an appropriate grid size (Section 3.5) and of a proper temporal length of the transient HSRs (Section 3.6). This last analysis is needed since, theoretically, the HSRs should store the data referring to the evolution of the system until steady state but, in practical situations, this means an unnecessarily high computational time. In Section 3.8, the steady state and transient FTMs for the finite dimensional stack configuration are validated with respect to the results obtained by analogous FEM models.

3.2 Lateral boundary conditions

The lateral sides of the die stack are assumed to be insulated. This condition is widely used in literature since the vertical faces of the chip are much smaller than the horizontal ones and the cooling solutions (heat sink, heat spreader, PCB, . . .) are applied on top and bottom of the stack. This means that the heat flow through the lateral boundaries towards the ambient is negligible [24, 37] and, therefore, it is considered to be zero. For more details about the validity of this hypothesis, please refer to Figure 5.1 and Section 5.2.

The mathematical expression of the PDEs governing the evolution of $\theta_{z_i}(x, y, z_j, t)$ in case of a stack configuration of finite dimension can be written as

$$\rho(z)c(z)\frac{\partial\theta_{z_i}(x,y,z_j,t)}{\partial t} - \nabla\cdot[k(z)\nabla\theta_{z_i}(x,y,z_j,t)] = q(x,y,z_i,t), \qquad (x,y,t)\in\Omega\times\mathbb{R}^+ \tag{3.1a}$$

$$k(z)\frac{\partial\theta_{z_i}(x,y,z,t)}{\partial z} = h_t\theta_{z_i}(x,y,z,t), \qquad z = z_t \tag{3.1b}$$

$$k(z)\frac{\partial\theta_{z_i}(x,y,z,t)}{\partial z} = -h_b\theta_{z_i}(x,y,z,t), \qquad z = z_b \tag{3.1c}$$

$$\frac{\partial\theta_{z_i}(x,y,z_j,t)}{\partial \boldsymbol{n}} = 0, \qquad (x,y,t)\in\partial\Omega\times\mathbb{R}^+ \tag{3.1d}$$

$$\theta_{z_i}(x,y,z_j,0) = 0 \qquad (x,y)\in\Omega. \tag{3.1e}$$

where Ω is the 2D spatial geometry in which $\theta_{z_i} = (x, y, z_j, t)$ is defined $\forall t$ and $\boldsymbol{n}$ represents the direction perpendicular to its boundary $\partial\Omega$. The fundamental idea behind the method of images, which was originally developed in electrostatics, is to create zero heat flux through the adiabatic boundaries by means of appropriately placed additional heat sources.

3.3 Method of images

In this Section the method of images is explained. Its graphical and intuitive illustration is presented in Section 3.3.1 for a semi-infinite case, while the corresponding rigorous mathematical derivation is reported in Section 3.3.2. Later on, in Section 3.3.3, the case for two dimensional, finite structures is considered.

3.3.1 Illustration: semi-infinite structure

The method of images is illustrated in Figure 3.2 for hot spot power dissipation in a simple, semi-infinite, 1D-structure. On the top line a schematic of the modeled geometry is shown: it is assumed to have an adiabatic boundary condition on

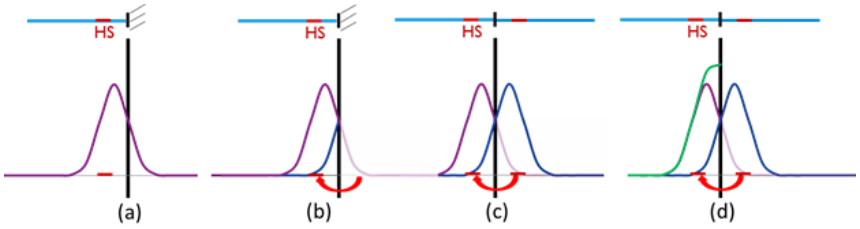


Figure 3.2: Method of images technique for a 1D semi-infinite domain.

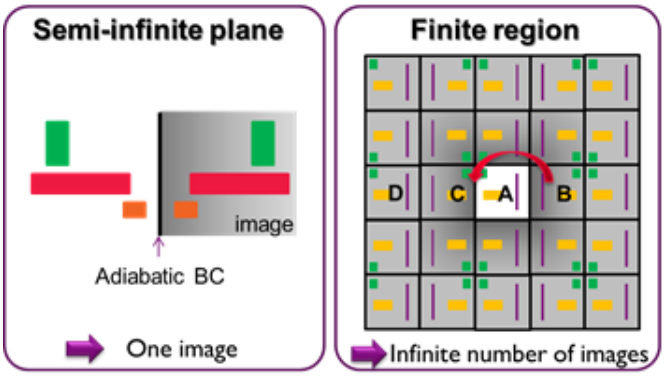


Figure 3.3: Method of images concept in case of a half-plane domain (left) and of a finite dimensional region (right) in 2D.

the right side and to be infinite on the left one. Hot spot power is dissipated in the region indicated by "HS". According to the superposition principle, the HSR, obtained considering an infinite 1D-structure, is centered over the power dissipation area and multiplied by the dissipated power density (a). The effect of the insulating boundary on the right hand side is the *reflection* of the heat flux (b). This effect can be easily modeled by placing a mirrored artificial heat source, symmetric with respect to the boundary, and the corresponding temperature response (c). Superposing these two HSRs the effect of the insulated boundary is captured (d). A similar approach is used for two dimensional PMs where, in case of a semi-infinite structure, the PM fills a half-plane. To model the adiabatic boundary condition, the original PM is flipped, using the insulated boundary as a mirror, and reported on the other half-plane (left hand side of Figure 3.3).

3.3.2 Mathematical derivation: semi-infinite structure

In this Section, the validity of the method of images as a way to model the insulating BC on a semi-infinite structure is mathematically proven. Let's consider the heat

conduction PDE in the half-plane $\Omega = \mathbb{R}^- \times \mathbb{R}$ with insulating BC and zero initial temperature

$$\rho(z)c(z)\frac{\partial \theta_{z_i}(x,y,z_j,t)}{\partial t} - \nabla \cdot [k(z)\nabla \theta_{z_i}(x,y,z_j,t)] = q(x,y,z_i,t), \qquad (x,y,t) \in \Omega \times \mathbb{R}^+ \tag{3.2a}$$

$$\frac{\partial \theta_{z_i}(x,y,z_j,t)}{\partial x} = 0, \qquad x = 0 \tag{3.2b}$$

$$\theta_{z_i}(x,y,z_j,0) = 0 \qquad (x,y) \in \Omega. \tag{3.2c}$$

Solving this PDE for θ_{z_i} by means of the method of images means, first of all, to extend the domain Ω over the whole space $\mathbb{R}^2$ and to solve the following PDE defined for $\theta^e_{z_i}$

$$\rho(z)c(z)\frac{\partial \theta^e_{z_i}(x,y,z_j,t)}{\partial t} - \nabla \cdot [k(z)\nabla \theta^e_{z_i}(x,y,z_j,t)] = q^{even}(x,y,z_i,t), \qquad (x,y,t) \in \mathbb{R}^2 \times \mathbb{R}^+ \tag{3.3a}$$

$$\theta^e_{z_i}(x,y,z_j,0) = 0 \qquad (x,y) \in \mathbb{R}^2. \tag{3.3b}$$

where

$$q^{even}(x,y,z_i,t) = \begin{cases} q(x,y,z_i,t), & \text{if } x \le 0 \\ q(-x,y,z_i,t), & \text{if } x > 0 \end{cases} \tag{3.4}$$

is the even extension of q over $\mathbb{R}^2 \times \mathbb{R}^+$.

What we want to prove is the equivalence between these two sets of equations. The first step consists in showing that:

$$\begin{aligned} q^{even}(x,y,z_i,t) \ \textit{even function} \Rightarrow \theta^e_{z_i}(x,y,z_j,t) \ \textit{even function} \\ \Leftrightarrow \theta^e_{z_i}(x,y,z_j,t) = \theta^e_{z_i}(-x,y,z_j,t). \end{aligned}$$

From equation (3.3a), performing a change of variable $-x \to s$ and exploiting the fact that q^{even} is an even function,

$$\begin{aligned} &\rho(z)c(z)\frac{\partial \theta^e_{z_i}(x,y,z_j,t)}{\partial t} - \nabla \cdot [k(z)\nabla \theta^e_{z_i}(x,y,z_j,t)] = q^{even}(x,y,z_i,t) \overset{s=-x}{\Rightarrow} \\ &\rho(z)c(z)\frac{\partial \theta^e_{z_i}(-s,y,z_j,t)}{\partial t} - \nabla \cdot [k(z)\nabla \theta^e_{z_i}(-s,y,z_j,t)] = q^{even}(-s,y,z_i,t) \\ &= q^{even}(s,y,z_i,t), \end{aligned} \tag{3.5}$$

the identity

$$\theta^e_{z_i}(x,y,z_j,t) = \theta^e_{z_i}(-x,y,z_j,t) \tag{3.6}$$

is obtained. Being $\theta^e_{z_i}(x,y,z_j,t)$ an even function with respect to the x variable, the following identity for the derivative with respect to x holds

$$\frac{\partial \theta^e_{z_i}(x,y,z_j,t)}{\partial x} = -\frac{\partial \theta^e_{z_i}(-x,y,z_j,t)}{\partial x}. \tag{3.7}$$

Together with the fact that $\theta^e_{z_i} \in C^2$, since it satisfies a second order PDE, the equality

$$\left.\frac{\partial \theta^e_{z_i}(x, y, z_j, t)}{\partial x}\right|_{x=0} = 0 \tag{3.8}$$

holds and, therefore, $\theta^e_{z_i}(x, y, z_j, t)$ satisfies the BC in equation (3.2b).

What is still missing to be proven, is that the restriction of $\theta^e_{z_i}$ to $\Omega = \mathbb{R}^- \times \mathbb{R}$ is also a solution of equation (3.2a). Since

$$\begin{aligned} &\rho(z)c(z)\left.\frac{\partial \theta^e_{z_i}(x, y, z_j, t)}{\partial t}\right|_{x\leq 0} - \boldsymbol{\nabla}\cdot[k(z)\boldsymbol{\nabla}\theta^e_{z_i}(x, y, z_j, t)|_{x\leq 0}] = q^{even}(x, y, z_i, t)|_{x\leq 0} \\ &= q(x, y, z_i, t) = \rho(z)c(z)\frac{\partial \theta_{z_i}(x, y, z_j, t)}{\partial t} - \boldsymbol{\nabla}\cdot[k(z)\boldsymbol{\nabla}\theta_{z_i}(x, y, z_j, t)], \end{aligned} \tag{3.9}$$

the mathematical validity of the method of images to solve PDEs over a half-plane with an insulating BC is proven.

3.3.3 Illustration: finite dimensional structure

The theory illustrated up to here is valid for semi-infinite structures in which just one insulating boundary is present. For a finite region with multiple adiabatic BCs, the zero-flux condition has to be fulfilled for all these boundaries. However, this condition is never satisfied exactly. Let's consider a function $q(x, y, z_j)$ defined over a square or a rectangle in the (x, y) plane, which represents the PM dissipated on a certain level z_j. This is the area indicated by A and with a white background color in Figure 3.3. Following the reasoning described for half-plane domains, to model the impact of insulating BCs, a ring of mirrored images should be added all around the original PM. However, as shown in Figure 3.3, adding a ring of images compensates for the heat flux coming from the central part but not for the one coming from the other added images. More precisely, adding image B, for example, compensates, on the one hand, for the heat going through the boundary $A - B$ but it also generates more heat flux that goes through $A - C$. In order to compensate for this flux, block D needs to be added. With analogous reasoning, an infinite number of images should be used to correctly model zero heat flux through the four boundaries.

Luckily, since our interest is Θ_{z_i}, which is defined in the area of the original PM, and since far enough from the dissipation point the heat flux is negligible, the farther away an image is from the original PM, the less its thermal impact on Θ_{z_i}. This means that considering an appropriate limited number of images results in a negligible flux through the boundaries and, therefore, in a negligible error. The HSRs matrices are, consequently, taken large enough to cover the extended power

Figure 3.4: Relationship between NI, the percentage relative error and the computational time.

maps (including images), PM^e, but the extra peripheral area is neglected avoiding useless computational costs.

The temperature increase $\Theta_{z_i}(\cdot,\cdot,z_j,\cdot)$ on layer z_j due to power dissipated on level z_i can, therefore, be obtained from the results of the convolution between the extended version of PM_i, PM_i^e, and $HSR_{z_i}(\cdot,\cdot,z_j,\cdot)$

$$\Theta^e_{z_i}(\cdot,\cdot,z_j,\cdot) = HSR_{z_i}(\cdot,\cdot,z_j,\cdot) * PM^e_{z_i} \tag{3.10}$$

where $*$ indicates 2D-convolution for steady state simulations and 3D-convolution for transient ones. The so obtained $\Theta^e_{z_i}(\cdot,\cdot,z_j,\cdot)$ refers to an area as large as PM_i^e: $\Theta_{z_i}(\cdot,\cdot,z_j,\cdot)$ is its central part, the part referring to the position of the original PM_i in the PM_i^e extended matrix.

The temperature increases $\Theta_{z_i}(\cdot,\cdot,z_j,\cdot)$ obtained in this way, are the discrete solutions of equations (3.1a)-(3.1e). The restriction of $\Theta^e_{z_i}(\cdot,\cdot,z_j,\cdot)$ on the finite domain Ω solves, indeed, equation (3.1a). The method of images takes care of the Neumann BC on the lateral boundaries of the stack (equation (3.1d)) while the impact of the top and bottom BCs (equations (3.1b)-(3.1c)) as well as of the initial zero temperature (equation (3.1e)) is included in the HSRs.

3.4 Required number of images

The selection of an appropriate number of images per side, NI ($NI = 2$ on the right hand side of Figure 3.3), is important to keep the computational time as low as possible while properly and accurately modeling the insulating BC. Figure 3.4 shows an example on how the percentage relative error, with respect to FEM, and the computational time vary as a function of NI. A uniform power map is

considered in this example and the percentage errors are computed comparing the maximum temperature increases achieved by using a FEM model and the corresponding FTM with different numbers of images. The use of five images instead of one results, in this case, in 900 times error reduction as well as in three times computational time increase. A further increase in *NI* mainly results in an increment in computational time without any significant accuracy improvement.

Given a certain desired accuracy level for the FTM solutions, with respect to FEM results of analogous structures, *NI* mainly depends on the width of the HSRs. Indeed, the narrower the HSR is, the smaller the region around the heat source where the temperature increases and the smaller the effect of the mirrored heat sources. Consequently, the number of required images, *NI*, will be lower for narrower HSRs. The width of the HSR depends on how the heat dissipates and spreads out in the stack configuration and this depends, in turn, in a complex way on the system parameters. This means that, given a certain required accuracy level $1 - \tilde{\alpha}$, it is not possible to extract a single value for *NI* that is valid for every situation. It is, however, possible to define the worst PM case, the PM for which, given a particular set of parameters defining the geometry, the materials and the BCs, the highest *NI* is needed to achieve that accuracy. If the *NI* corresponding to this *PM* and to a specific $\tilde{\alpha}$ value is selected, we are sure that the effect of the insulating BCs is included with a relative error less than $\tilde{\alpha}$. After several tests, this PM has been found to be the case of uniform dissipation. Moreover, it is enough to determine *NI* in the steady state regime since this is the regime in which the heat reaches its maximum lateral spreading. In other words, this is the regime in which the HSRs reach their maximum width and, as a consequence, it's the situation in which the higher *NI* is needed to ensure a specific accuracy.

It also important to stress that, when we talk about *accuracy*, the results of analogous models solved by FTM and by FEM are compared. In this Chapter, for example, it means that the same geometry, constituted by multiple stacked layers and similar to the sketch in Figure 3.1, the same material properties and the same BCs are considered in the two models. The comparison of analogous models allows to check, step by step, the accuracy of the solutions implemented in the FTM to overcome each particular limitation. If, on the contrary, the FEM results for the complete package structure would have been compared with the actual FTM output, the overall accuracy of the FTM at *that* stage of development could have been checked, not the one of the specific implemented solution (how to model insulating BCs in this case).

3.4.1 Temperature computation for uniform power dissipation: *annulus method*

As already mentioned, the worst PM case, in terms of number of images needed to accurately model the insulating BCs, is uniform power dissipation in steady

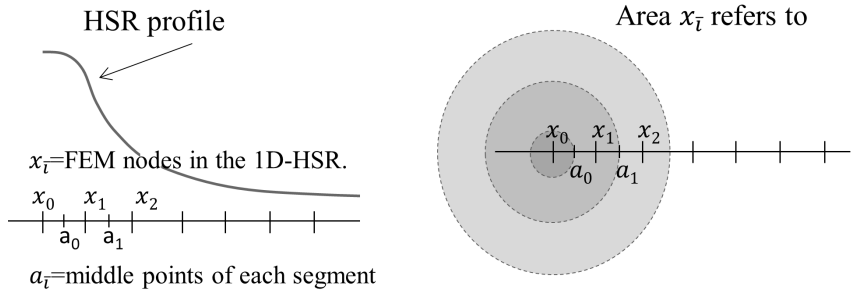


Figure 3.5: Illustration of the *annulus method*, i.e. the fast methodology developed to compute the temperature due to a uniform PM.

state. In order to develop a methodology able to predict, for each specific case, an appropriate *NI*, an algorithm is proposed to quickly compute the temperature increase due to uniform power dissipation. This can be done implementing a resistance network but, since all the information related to the thermal behavior of the system is stored in the HSRs, which need to be calculated anyhow for the FTM, we will use them to reach our objective.

The *annulus method* is basically a simplification of the 2D-convolution, valid in case one of the two matrices (the PM in this case) is uniform. Under this circumstance, and for the stack configuration, indeed, the temperature increase is uniform and each value of the matrix resulting from the convolution, can be computed as

$$\Theta_{z_i}(\cdot,\cdot,z_j) = \sum_{\bar{i},\bar{j}} PM_{z_i} \cdot HSR_{z_i}(\bar{i},\bar{j},z_j)$$

where $\bar{i}, \bar{j}$ are row and column indexes. However, since the HSR has circular symmetry, the calculations can be further simplified considering a 1D-HSR vector, with data as a function of the distance from the HS (cf. Figure 3.5). This vector can be easily obtained by 1D-interpolation and normalization of the results from the axisymmetric FEM. By using the 1D-HSR vector, however, care should be taken on how many terms in the sum refer to the same value in the 1D-HSR vector. In other words, we need to know how much area of the original PM refers to each single 1D-HSR value. This can be performed, once more, exploiting the circular symmetry property, considering the values in the PM in W/m^2 units, and multiplying each term in the sum by the corresponding circular annulus area:

$$\Theta_{z_i}(\cdot,\cdot,z_j) = \frac{PM_{z_i}}{\bar{h}^2}[\text{1D-}HSR(0,z_j)]\pi a_0^2 + \sum_{\bar{i}\geq 0} \frac{PM_{z_i}}{\bar{h}^2}[\text{1D-}HSR(\bar{i},z_j)]\pi(a_{\bar{i}+1} - a_{\bar{i}})^2 \quad (3.11)$$

where $a_{\bar{i}}$ is the middle point between the locations 1D-$HSR(\bar{i},z_j)$ and 1D-$HSR(\bar{i}+1,z_j)$ refer to, and $\bar{h}^2$ is the area of one cell in the PM.

3.4.2 Method to predict the number of images

In this Section a method is presented to compute NI, which is the minimum number of images per side of the PM, so that the error due to the modeling of the insulating boundary conditions is lower than a user defined quantity $\tilde{\alpha}$ in the worst PM case scenario. This operation can be quickly performed in the preprocessing phase, after the 1D-HSRs have been extracted from FEM, by means of the *annulus method*.

Θ_{z_i} in equation (3.11) is the uniform temperature obtained considering the data in the full length of the 1D-HSR. If the summation in the same equation is run considering just the first $\bar{n}$ values in the HSRs, then the obtained temperature increase

$$\Theta^{\bar{n}}_{z_i}(\cdot,\cdot,z_j) = \frac{PM_{z_i}}{\bar{h}^2}[\text{1D-}HSR(0,z_j)]\pi a_0^2 + \sum_{\bar{i}=0}^{\bar{n}-1} \frac{PM_{z_i}}{\bar{h}^2}[\text{1D-}HSR(\bar{i},z_j)]\pi(a_{\bar{i}+1}-a_{\bar{i}})^2, \quad (3.12)$$

is computed assuming that just the distance up to $x_{\bar{n}}$ is covered by images. We can now define the relative errors between $\Theta^{\bar{n}}_{z_i}$ and Θ_{z_i}, which is considered as the correct solution, as

$$err_{\bar{n}} = \frac{\Theta_{z_i} - \Theta^{\bar{n}}_{z_i}}{\Theta_{z_i}}. \quad (3.13)$$

The minimum value of $\bar{n}$ for which $err_{\bar{n}} < \tilde{\alpha}$ is then selected. From this value, $\bar{n}^*$, the minimum distance that has to be covered by images from the boundary of the PM can be recovered. However, this is a circular method while PM^e covers a square area. For this reason, a circle to square transformation is performed by trying to maintain the measure of the area. However, NI can just assume integer values. As a consequence, NI is chosen as the minimum value so that the required specified area is covered:

$$NI = \left\lceil \frac{\sqrt{\pi} a_{\bar{n}^*}}{2\, cs} \right\rceil \quad (3.14)$$

where cs is the chip size and $\lceil x \rceil$ indicates the smallest integer number greater than or equal to x.

This algorithm has been tested for different $\tilde{\alpha}$ values (0.01, 0.03 and 0.05) and for 50 different geometries. The tested structure was constituted by just one block that, in different experiments, could assume different conductivity values and thicknesses. Insulation was imposed on all the boundaries except for the bottom one where convection was applied. The choice of letting the thermal conductivity of the block vary relies on to the fact that, in a more realistic design, more chips are stacked on top of each other and layers of material with low thermal conductivity are placed between them. Thus, the different values may be interpreted as equivalent conductivity values in case of more stacked dies, different underfill materials and different interfaces/dies thicknesses. Even if the equivalent conductivity of the

Parameter	Min	Max
cs (m)	0.003	0.02
h_b (W/m^2K)	700	1500
Power density (W/m^2)	5000	15000
Chip thickness (μm)	50	1000
Thermal conductivity (W/mK)	40	150

Table 3.1: Parameters used to check the validity of the algorithm proposed to define NI. Values are chosen randomly between the minimum and maximum.

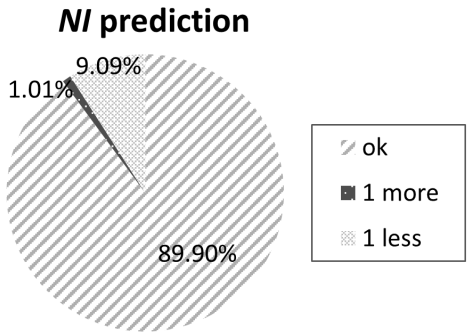


Figure 3.6: Reliability of the method to predict NI: graph illustrating the percentage with which the method predicts the correct number of images, one more or one less, with respect to the real number needed to achieve a certain accuracy $1 - \tilde{\alpha}$.

stack should be orthotropic ($k_x = k_y \neq k_z$), since uniform power dissipation is considered and, as a consequence, there is no lateral spreading, the use of k_z as isotropic value does not affect the temperature result. The values of the parameters describing the geometry and the material property have been chosen randomly between the limits reported in Table 3.1.

The results of these tests are shown in Figure 3.6. The data are obtained by comparing the value of NI, estimated by the proposed algorithm, to the smallest number of images needed in the convolution based FTM to have an error less or equal to $\tilde{\alpha}$, when compared to analogous FEM. As the pie plot shows, this algorithm correctly predicts NI in almost 90% of the cases. In 1% of the cases one image more is forecast, resulting in a higher computational time and a higher accuracy than required, while in 9% of the cases one image less is predicted causing an error higher than the desired $\tilde{\alpha}$ in the worst PM case.

Once NI has been defined, the optimal number of elements of the HSRs can be determined. This is necessary because, although the solution of the 2D-axisymmetric FEM model is quickly obtained even for a large structure, care should be taken on the size of the convolved elements in the FTM to keep

the computational time as low as possible. If, for example, the 2D- or 3D-HSRs are computed starting from N elements in the 1D-HSRs instead of $N-1$, $[N+(N-1)]^2-[(N-1)+(N-2)]^2 = 8(N-1)$ more elements are considered in the 2D-HSRs and in each time layer of the 3D-HSRs. The optimal size of the HSRs depends on the dimensions of PM^e and it is taken so that, when computing $T_{z_i}(\bar{i},\bar{j},z_j)$ in equation (2.29), $PM^e_{z_i}(\bar{i}-m,\bar{j}-n)$ is defined $\forall m,n$. Since the peak of the HSRs has to be in the center of the 2D-HSRs matrices, the optimal dimensions of the HSRs, based on the selected NI value, are $\lceil(2\cdot NI\cdot cs)/\bar{h}+1\rceil\times\lceil(2\cdot NI\cdot cs)/\bar{h}+1\rceil$.

While dealing with FTM of 3D-ICs, square footprints are normally considered for the die stack. However, this FTM is also applicable to rectangular shapes. The only difference is that, for rectangular footprints, the value of NI depends on the considered edge of the stack. NI is initially computed, as explained in this Section, for the longest side of the rectangle. NI on the shortest edge is, then, calculated so that at least the same distance outside the chip, as on the long side, is covered by images.

3.4.3 Algorithm

The algorithm for the computation of the number of images, NI, required to ensure a certain accuracy level, $1-\tilde{\alpha}$, of the FTM is summarized in the flowchart in Figure 3.7. The *annulus method*, which is the algorithm concerning how to compute the temperature increase due to uniform power dissipation starting from the corresponding HSR, is highlighted in the chart.

3.5 Spatial grid size

Another parameter that affects the accuracy of the FTM results is the selection of the resolution of the extracted HSRs, i.e. the selection of the value $\bar{h}$ of the grid size for the discretization. This parameter, indeed, affects the spatial accuracy and plays a role both in steady state and transient simulations. In principle, one can assume a grid size $\bar{h}$ as large as the smallest hot spot in the applied PMs. However, this approach can cause large inaccuracy and, therefore, a smaller value of $\bar{h}$ may be desirable. As for all the problems solved in discrete domains, the smaller the grid size, the higher the accuracy. However, this comes at the expenses of the computational time and, as a consequence, a trade-off between these two quantities has to be considered.

Figure 3.8 shows, for a particular case, the relationship between the mesh size $\bar{h}$, the relative error with respect to the FEM results in the location of the maximum temperature (bars) and the computational time required by the FTM (curve). The considered power map presents three hot spots of $50\times50\ \mu m^2$, three of $150\times150\ \mu m^2$

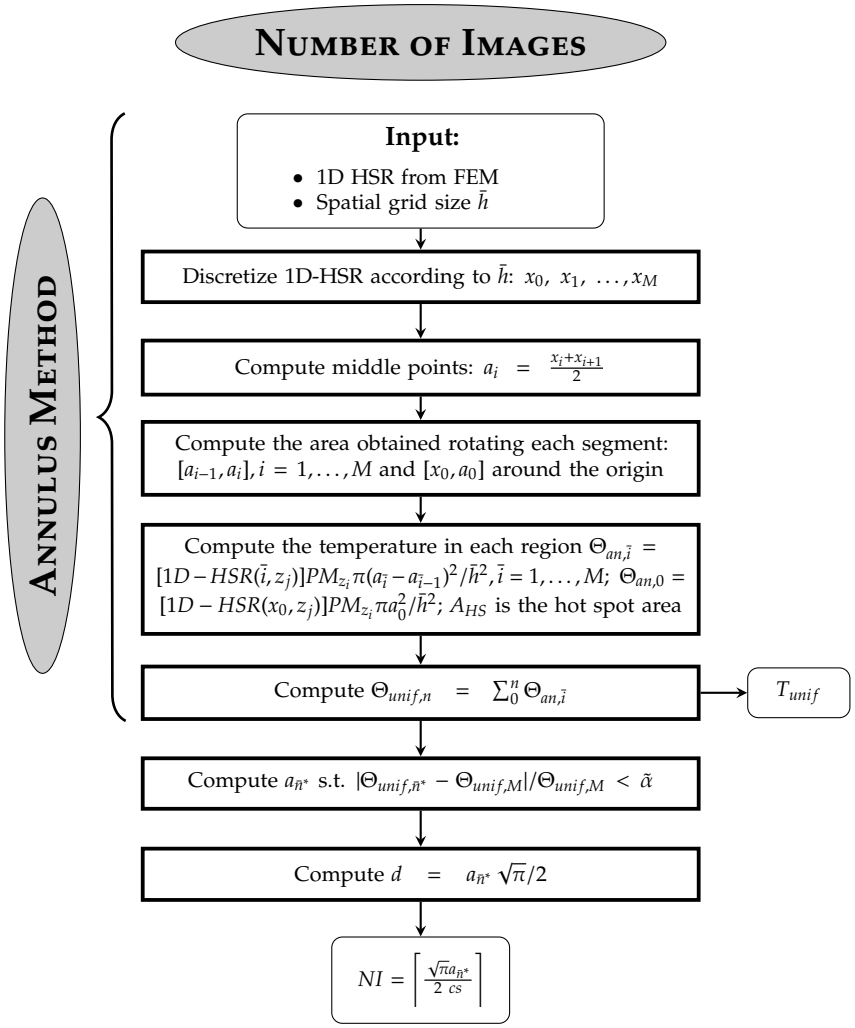


Figure 3.7: Flowchart representing the algorithm of the *annulus method* and the algorithm implemented to compute the number of images NI.

and three of $500 \times 500\ \mu m^2$. In each of them 0.1 W is dissipated and the dimensions of the chip are $10.5 \times 10.5\ cm^2$. Even if the algorithm to determine the number of images returns $NI = 4$ in case of $\bar{h} = 50\ \mu m$ and $\tilde{\alpha} = 5\%$, since in this example the PM is fixed, it is highly non-uniform and with a HS far from the chip edge, the value $NI = 1$ has been proven to be enough to accurately model the temperature profiles in this specific case. It has, therefore, been used to obtain the data in Figure 3.8. Detailed information about the fabrication of the considered structure can be found in [77]. The graph highlights that the selection of a proper value of $\bar{h}$ is

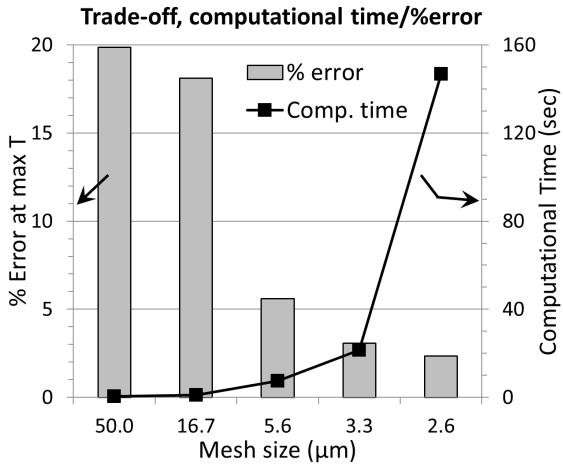


Figure 3.8: Relationship between the mesh size, the relative %error with respect to the FEM results in the location of the maximum temperature for a HS of 50 μm (bars, left vertical axis) and the computational time (curve, right vertical axis).

essential to ensure good accuracy of the results. For this specific case, for example, a reduction of the mesh size by a factor of 19, starting from 50 μm, results in a decrease of more than eight times in the relative percentage error. However, since the computational time behaves as $O(N \log N)$, with N the number of elements in the extended matrices (cf. Section 2.5.3), this accuracy improvement is associated with a 300 times increase in computational time. Moreover, looking at the graph, a clear reduction of the error is evident up to $\bar{h} = 3.3\ \mu m$; a further reduction of $\bar{h}$ is associated with a limited improvement in accuracy and with a high increase in computational time. This confirms the necessity of considering a trade-off between computational time and accuracy.

The main reason underlying this accuracy issue is the discretization of the HSR. During this process, indeed, a *single* temperature value is assigned to each cell used to discretize this function. This means that, the faster the *continuous* HSR varies within a cell, the higher the corresponding discretization error is. For this reason, given a particular mesh size $\bar{h}$, the magnitude of the relative error with respect to the FEM solution and the reduction of this error, achievable by using a smaller value of $\bar{h}$, can't be established a priori but they are case dependent. Changing the power maps, the boundary conditions, the material properties and the geometry of the modeled structure, different values of these two quantities are obtained. Similarly to the approach described for the calculation of NI, the worst PM case has been considered. For this analysis, in particular, it is the one that, given a specific structure, requires the finest resolution to achieve a specific accuracy. While in case of the definition of NI, the uniform PM represents the worst case scenario, in case of the calculation of $\bar{h}$, the worst PM case is the one

in which power is dissipated in hot spots. Contrary to the *NI* case, when talking about the selection of an appropriate value of $\bar{h}$, the worst PM is not 100% fixed: the dimension of the selected HS can, indeed, vary. In practice, the HS can be considered as large as the smallest power dissipation area that can be possibly selected for that particular structure.

In order to define an appropriate value of $\bar{h}$, the maximum temperature, due to a hot spot power dissipation of size $\bar{h}^{HS}$, is computed by the FTM for different values of $\bar{h}$. More precisely, the result obtained considering a grid size $\bar{h}_k$ is compared to the one obtained using a grid size $\bar{h}_{k+1} = \bar{h}_k/d$, with d an odd number ($d = 3, 9, 15, 19$ in Figure 3.8). This last requirement is related to the fact that the temperature values computed by the FTM refer to the center of each cell: if $\bar{h}_k$ would be divided by an even number, the locations where the temperature is computed change. The improvement in accuracy using d^2 times more elements in the PM and in the HSR can, then, be related to the increase in computational time and a trade-off can be defined. To allow for a fast determination of this trade-off, a dedicated algorithm to compute the maximum temperature increase, $\max(\Theta^k)$, due to hot spot power dissipation and a specific grid size $\bar{h}_k$, has been developed and tested.

The implemented methodology resembles the *annulus method* and has been developed for the steady state regime. The proposed algorithm is, indeed, a fast implementation of the convolution idea, valid if the power dissipated in the PM is restricted to just *one* square area with an edge size of $\bar{h}^{HS}$ and if the power density within this area is constant. The discrete equation allowing the calculation of the temperature increase in a specific point (cf. Section 2.5.3),

$$\Theta_{z_i;\bar{h}_k}(\bar{i}, \bar{j}, z_i) = \sum_{m=-a}^{a} \sum_{n=-b}^{b} HSR_{z_i}(m, n, z_i) PM_{z_i}(\bar{i} - m, \bar{j} - n), \tag{3.15}$$

can, indeed, be adapted to this specific power map. Since just a limited number of cells in the considered PM are active, the summations in equation (3.15) can be reduced in such a way that just these *active* cells are included. If $\bar{h}_k$ indicates the considered grid size and $\bar{h}^{HS}$ indicates the size of the HS, then the number of active cells in the PM surrounding, in each direction, the center of the HS can be computed as $l = \left\lfloor \frac{\bar{h}^{HS}}{2\bar{h}_k} \right\rfloor$ ($\lfloor x \rfloor$ indicates the greatest integer number smaller than or equal to x). Moreover, since all the active cells in this PM have the same value, by assuming the center of the HS in position $(\bar{i}, \bar{j}, z_i)$, equation (3.15) reduces to

$$\max(\Theta_{\bar{h}_k}) = \Theta_{z_i;\bar{h}_k}(\bar{i}, \bar{j}, z_i) = PM_{z_i}(\bar{i}, \bar{j}) \sum_{m=-l}^{l} \sum_{n=-l}^{l} HSR_{z_i}(m, n, z_i). \tag{3.16}$$

In order to obtain an estimation of the error due to a certain discretization of a PM with a HS of size $\bar{h}^{HS}$, it is enough to compare the values obtained by equation (3.16) considering the HSR and the PM referring to different grid sizes. More precisely, if the first attempt in computing $\max(\Theta_{\bar{h}_k})$ is obtained considering a grid

of size $\bar{h}_k$, after having adapted the PM and the HSR for a grid size of $\bar{h}_{k+1} = \frac{\bar{h}_k}{d}$, with d and odd number, $\max(\Theta_{\bar{h}_{k+1}})$ can be easily obtained from equation (3.16). It is worth to remind at this point that, if the value of $\bar{h}$ changes, then, according to the discussion in Section 2.5.1, the HSRs need to be recomputed adapting the size of the generating HS. By reducing the grid size, the error tends to zero. Thus, the relative difference between the values $\max(\Theta_{\bar{h}_k})$ and $\max(\Theta_{\bar{h}_{k+1}})$ provides an estimation of the inaccuracy of the numerical evaluation. By repeating this step multiple times and by knowing that the computational time of the convolution algorithm goes as $O(N \log N)$, with N the number of elements in the extended matrices, it is possible to define a trade-off between the required accuracy and the computational time.

3.6 Time length of the HSR in transient regime

As mentioned in Section 2.5.1, for transient simulations, the FEM model from which the transient HSRs are extracted is run until steady state is reached. This is possible without high computational costs because, since the FEM works with an adaptive time step, large time steps can be used when the system is close to the steady state. In this situation, indeed, the temperature variation is slow. However, when the FTM is implemented, a fixed time step Δt needs to be chosen and the discretization of the HSRs, with that fixed Δt, until *complete* steady state could cause an increase in computational time that is not followed by a corresponding improvement in accuracy. The temperature response to an impulsive power dissipation presents, indeed, a high variation during the heating up and in the beginning of the cooling down phases but, afterwards, it slowly tends to zero. This means that, when performing superposition (or convolution) in time, the terms referring to the time steps in which the values of the HSR are close to zero (far away past) generate a negligible contribution to the overall temperature at present time. However, the impact on computational time may be significant because, for each further time layer considered in the 3D-HSR, two extra 2D-HSRs are included in the convolution. A trade-off between computational time and accuracy should, therefore, be selected. This means that a time step $\bar{t}_{ss}$ at which the HSRs have to be truncated needs to be defined and that $HSR(\cdot, \cdot, \cdot, \bar{t}) = 0, \ \forall \bar{t} \geq \bar{t}_{ss}$. However, as for the selection of NI and $\bar{h}$, an appropriate *truncation time* depends on all the design parameters and, as a consequence, a fixed value, independent of the considered configuration, cannot be defined. It is, however, possible to make use of a strategy, similar to the one implemented to calculate NI (cf. Section 3.4), to estimate the error introduced in the model by *truncating* the 3D-HSR at time step $\bar{t}_{ss}$.

Since the main interest is in obtaining accurate results in the locations of high temperature and since the scope of this Section is to estimate the error due to the truncation of the HSRs and not to the temporal/spatial variation of the PMs, the developed error estimation methodology is based on uniform power dissipation in

space, continuous in time. The input of the algorithm is the time dependent HSR extracted from the 2D-axisymmetric FEM model solved with adaptive time step. It is considered after time interpolation but before the mapping of the 1D spatial information at each time step into 2D matrices. The main steps of the algorithm are the following ones.

1. Compute the temperature increase due to uniform and impulsive power dissipation according to the values stored in each individual time layer, $\bar{t}_k$, of the HSRs. This is performed by multiple runs of the *annulus method* algorithm presented in Section 3.4.3. The input parameters of the algorithm are 1D-$HSR(\cdot,\cdot,z_j,\bar{t}_k)$ and $\bar{h}$ while the outputs are $\bar{\Theta}_{unif}(\cdot,\cdot,z_j,\bar{t}_k;\bar{t}_0)$. More precisely,

$$\bar{\Theta}_{unif}(\cdot,\cdot,z_j,\bar{t}_k;\bar{t}_0) = \sum_{dist}\Big[HSR(dist,z_j,\bar{t}_k)A_{dist}\Big]PM/\bar{h}^2 \tag{3.17}$$

   where A_{dist} is the spatial area the value $HSR(dist,z_j,\bar{t}_k)$ refers to (cf. Section 3.4.3) and PM represents the value stored in each cell of the uniform power map.

2. Compute $\Theta_{unif}(\bar{t}_k) = \sum_{i=0}^{k}\bar{\Theta}_{unif}(\bar{t}_i)$, which are the temperature increases due to uniform and continuous power dissipation in case the HSRs is truncated at time $\bar{t}_k$. This formula relies on the fact that the dissipated power is constant in time and that the power dissipated at $\bar{t}_i$, with $i \leq k$, affects the temperature in $\bar{t}_k$.

3. Compute the relative percentage error, $\%error = \frac{\Theta_{unif}(\bar{t}_{end})-\Theta_{unif}(\bar{t}_k)}{\Theta_{unif}(\bar{t}_{end})}$, where $\bar{t}_{end}$ is the last time step available in the HSR extracted from FEM, and select the time step $\bar{t}_{ss}$ for which the error is less than a user defined value $\tilde{\alpha}$. The percentage error is considered here, even if the transient regime is being analyzed, because just the heating up phase is modeled.

It is worth to note that the truncation time is strictly related to the time constant τ of the system.

Definition 5. The *time constant* τ is the parameter characterizing the response of the system to a step power dissipation. It is defined as the point in time at which the system's step response reaches the value of $1 - 1/e \approx 63.2\%$ of its steady state temperature increase.

In order to show the importance of selecting a proper value for $\bar{t}_{ss}$, three different cases have been considered and the results reported in Figure 3.9. The differences between the simulated structures are in the applied boundary conditions and in the inclusion of a plastic layer with high thermal capacitance on top of the stack. The related parameters are listed in Table 3.2. High capacitance, as in case 3, means

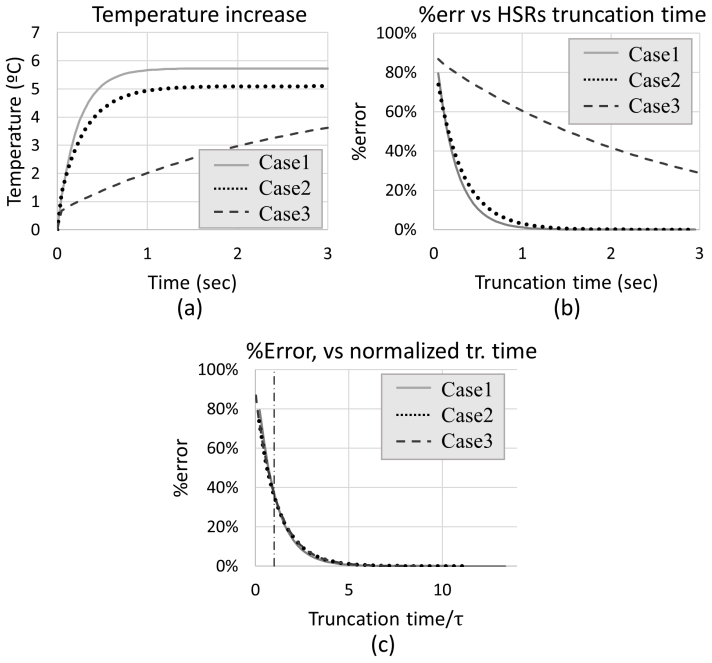


Figure 3.9: (a): Temperature increases as a function of time; (b): %errors as a function of the truncation time in the HSR; (c): %errors as a function of the normalized truncation time in the HSR. The results refer to the three cases listed in Table 3.2.

Parameters		Case 1	Case 2	Case 3
Plastic layer top	Thickness $[mm]$	-	1	1
	c $[J/kg\ K]$	-	11000	110000
	ρ $[kg/m^3]$	-	1150	1150
τ $[s]$		0.221	0.224	2.3412

Table 3.2: Parameters used to obtain the data in Figure 3.9.

a slower evolution of the system (dashed curve) and, as a consequence, a higher τ value and a larger number of time steps to be kept in the HSRs in order to obtain accurate enough temperature estimation. Figure 3.9 (a) shows the temperature responses of these three systems to a step power dissipation.

The %error, calculated as explained previously in this Section, due to the truncation of the transient HSRs at a certain time is shown in Figure 3.9 (b). The case with a really high capacitance (case 3, dashed line) shows, as expected, a much higher error than the other two cases when the HSRs are truncated at the same time.

Figure 3.9 (c) shows the same results after the normalization of the truncation time with respect to the time constant τ of each system. The time constants are computed considering the steady state temperature values obtained by the corresponding FEM models. Since all the curves are now really similar, they can be easily fitted by a single function, $\%error = 0.87\exp(-0.87t/\tau)$. This means that, for each considered $\tilde{\alpha}$ value, the normalized time at which the transient HSRs are truncated is always the same. The real truncation time, however, can significantly vary due to variation of the τ parameter.

3.7 Flowcharts of the FTM algorithms

In this Section, the flowcharts reporting all the steps of the algorithms developed for the steady state and for the transient FTM are presented, respectively, in Figures 3.10 and 3.11. Boxes with rounded corners are used for inputs and outputs while rectangles with thick borders indicate blocks in which computations are performed. A gray background is used to indicate that the computations are performed by FEM (Msc Marc [69]), while a white background that they are performed by Matlab [66]. With respect to the final flowcharts presented in the previous Chapter (Figures 2.11 and 2.12), the computation of NI and of $\bar{t}_{ss}$, as well as the extension of the PMs according to NI, are added in the algorithms.

3.8 FEM validation

3.8.1 Modeled geometry

The FTM has been validated, both for the steady state and the transient regime, by comparison with the results obtained by a general purpose finite element software [69]. The FEM solution is, therefore, considered as the *reference* solution. In both cases, exactly the same structure has been modeled (no package). A two die stack in a face-to-face configuration (power dissipated on the bottom of the top die and on top of the bottom die) is considered. Homogeneous interface material is assumed in between the two dies. The design parameters for this test case are listed in Table 3.3 while the FEM setup is shown in Figure 3.12. Exploiting results from the accuracy assessment, one image per side is included in the FTM ($\tilde{\alpha} = 0.005$) and a grid size of 120 μm is used [57].

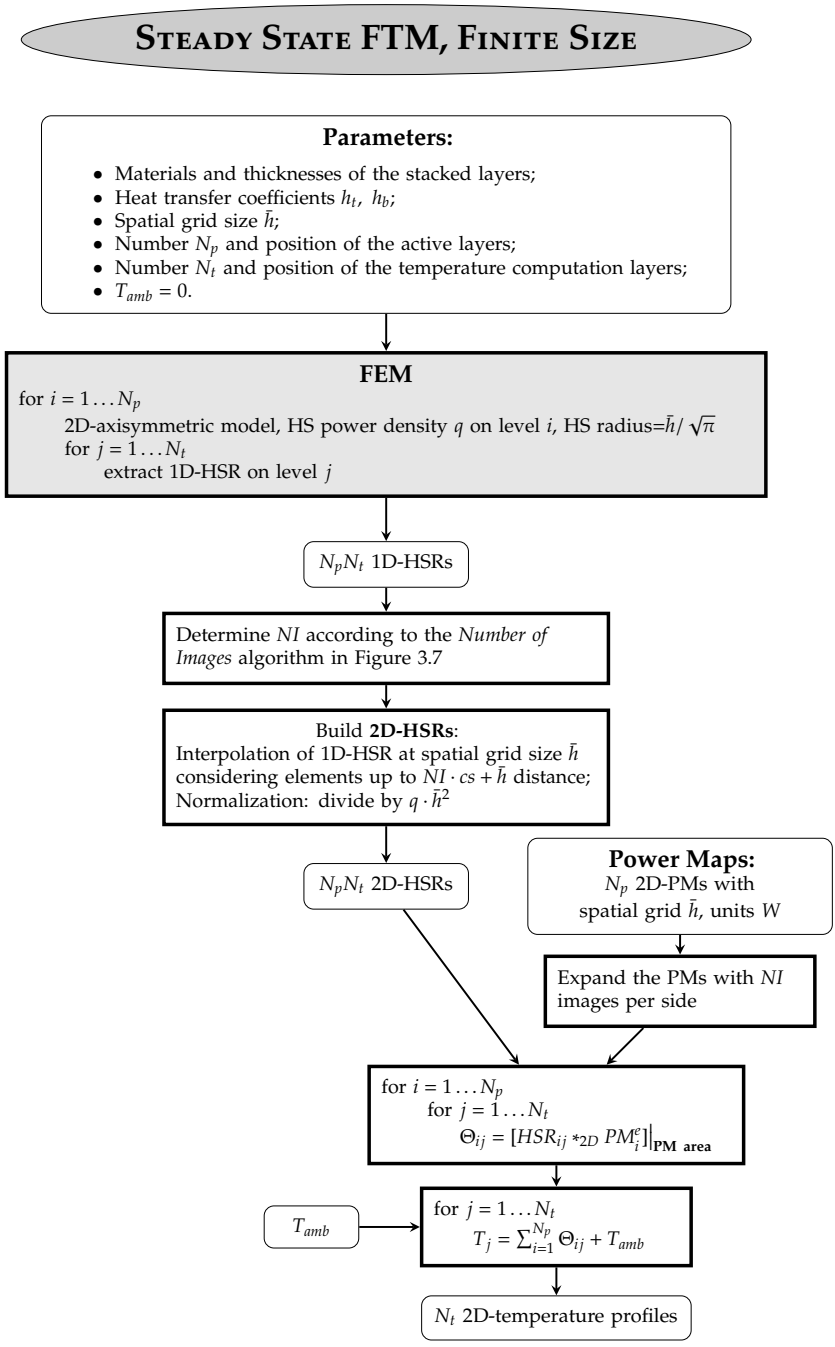


Figure 3.10: Flowchart representing the algorithm implemented for the steady state fast thermal modeling of 3D-stacks with finite horizontal size.

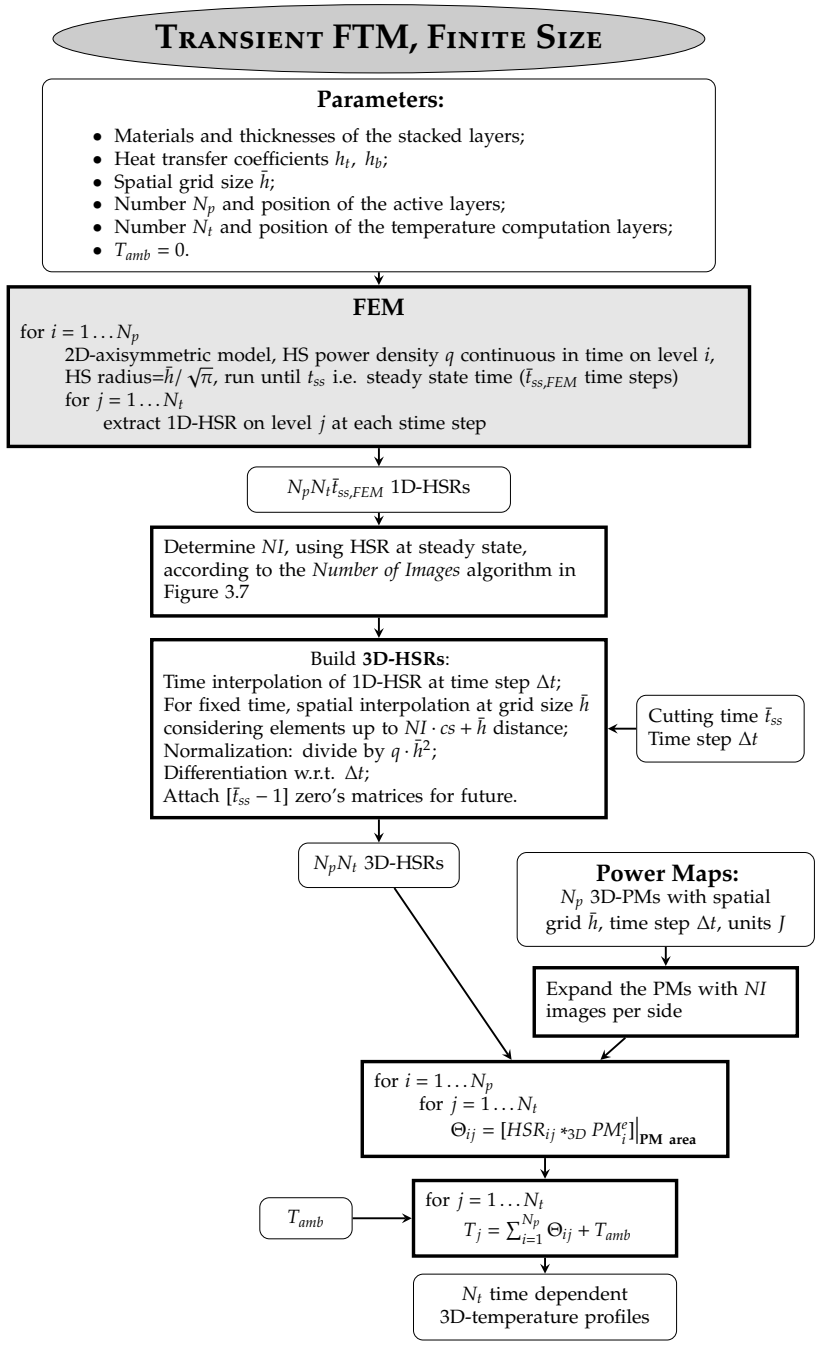


Figure 3.11: Flowchart representing the algorithm implemented for the transient fast thermal modeling of 3D-stacks with finite horizontal size.

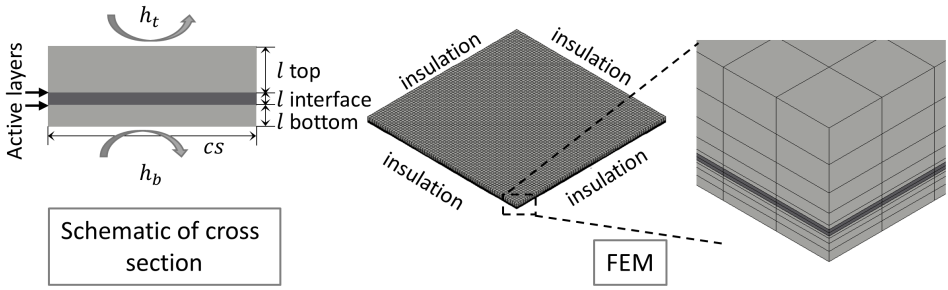


Figure 3.12: FEM setup used to validate the FTM for structures with finite size. The values of the parameters are reported in Table 3.3.

Geometry/material parameters			
Parameter	Value	Parameter	Value
cs	8.16 mm	l top die	200 μm
l interface	13 μm	l bottom die	50 m
k Si	120 W/mK	k interface	1 W/mK
c Si	700 J/kgK	ρ Si	2330 kg/m^3
c interface	2187 J/kgK	ρ interface	1051 kg/m^3

Boundary conditions			
Steady state		Transient	
Boundary	Value	Boundary	Value
Top	insulation	Top	$h_t = 20W/m^2K$ ($R_{th} = 780K/W$)
Bottom	$h_b = 1000W/m^2K$ ($R_{th} = 15.6K/W$)	Bottom	$h_b = 2000W/m^2K$ ($R_{th} = 7.8K/W$)
Lateral	insulation	Lateral	insulation

Table 3.3: Parameters and BCs used in the steady state and transient validations of FTM including the method of images.

3.8.2 Error metric

While dealing with accuracy, the metric of the error needs to be defined. Two different ways have been considered throughout this thesis: the percentage error, based on the temperature increase with respect to T_{amb},

$$\%err = \frac{|FTM - FEM|}{FEM - T_{amb}}, \tag{3.18}$$

and the absolute error,

$$|err| = |FTM - FEM|, \tag{3.19}$$

where *FTM* and *FEM* indicate, respectively, the solutions obtained by the FTM and FEM. After having checked the grid independence of the FEM results, the FEM solution is considered as the correct ones and it is used as the reference value for the estimation of the FTM error. Moreover, even if the full spatial map of the error (in both metrics) can be computed, sometimes, in the following of this thesis, also single numbers are reported. Whether they refer to the maximum error, the error in the location of the maximum temperature or the average error is specified according to each specific case.

The advantage of the percentage error is that it is independent of the dissipated power density. The disadvantage is that it is prone to assume high values in regions where the temperature is low, due to the low value of the denominator. This issue is exacerbated during cooling down phases in transient regime. Concerning $|err|$, the advantages and disadvantages are inverted: it depends, indeed, on the intensity of the dissipated power density but the maximum is assumed in the hot regions and during chip activity. These are the locations in which it is more relevant to have accurate results because thermally driven chip failures mainly occur at high temperatures.

For these reasons, $|err|$ is considered as error metric for transient simulations. However, since the cooling down phases are never included in steady state simulations, $\%err$ is used as error metric in this regime. The advantage of having a power density independent error is, indeed, predominant in this situation. Care should be taken in case of non-uniform PMs to locate max($\%err$) since it may occur in non-relevant, low temperature areas.

In the next Subsections these error metrics are used to evaluate the accuracy of the steady state and the transient FTMs for stack of dies of finite size with respect to FEM simulations of analogous structures.

3.8.3 Steady state regime

Results concerning the validation of the FTM with respect to an analogous FEM model are shown in Figure 3.13 for the steady state regime. Quantities referring to the top die are on the first row while the ones concerning the bottom die on the second row. The steady state PMs are shown on the first column and a total of 8.28 W is dissipated on the top die while 4.11 W on the bottom die. On the second column of the same Figure, the temperature profiles, calculated by the FTM on the heat dissipation levels, are presented. Insulation is assumed on all the boundaries except on the bottom side of the stack where convection, with $h_b = 1000\ W/m^2K$ (equivalent $R_{th} = 15.6\ K/W$), is applied. An ambient temperature of 25 $°C$ is considered. The percentage relative error with respect to FEM, which is

shown in the last column of the picture, is less than 1.12% and the improvement in computational time is more than 220 times (FTM 0.072 sec, FEM 15.93 sec). In both cases, just the solution time is considered, not the one that is needed to build the models and/or to obtain the input parameters.

The achieved accuracy is lower than the selected value of $\tilde{\alpha} = 0.5\%$ because the power is not uniformly dissipated: part of the error originates from the spreading happening within the die. Moreover, max(%*err*) is achieved in locations where power is not dissipated and where the temperature is low. In regions corresponding to power dissipation, the error is lower than 0.5%, in agreement with $\tilde{\alpha}$. This shows that the FTM is able to accurately model the steady state temperature profiles in 3D die stacks, significantly increasing the computational speed.

3.8.4 Transient regime

Constant power map

In the first case that is analyzed for the FEM validation in the transient regime, the behavior of the system is studied for constant power dissipation. This is to prove the ability of the FTM to detect the time dependent thermal evolution of the system. The same geometry used in the steady state simulation is considered. The only differences are in the ambient temperature and in the applied boundary conditions: in this case, convection is assumed both on the top and bottom surface of the stack. More precisely, $h_t = 20\ W/m^2$, $h_b = 2000\ W/m^2$ and $T_{amb} = 0\ °C$. The heating up process of the system is monitored for 1 sec, with a time step $\Delta t = 50\ msec$. The power is continuously dissipated, only on the bottom die, according to the bottom PM shown in Figure 3.13 (since this is a transient simulation, values are multiplied by Δt to represent the dissipated energy).

The results are presented in Figure 3.14 where the temperature increases on the top and bottom die are shown, respectively, on the first and the second row as a function of time. The Figure clearly shows that the calculated temperature maps are time dependent. To prove the ability of the FTM to detect the thermal transient behavior of the stack, the temporal evolution of the temperature increases obtained by the FTM is compared, in Figure 3.15 (a), with the corresponding results obtained by FEM. This is performed for the location corresponding to the hottest temperature on the top and bottom die. Curves refer to the FEM results while crosses to the FTM ones. Blue color is used for the bottom die while red for the top die. This Figure confirms the really good agreement between the two methodologies.

The high accuracy of this FTM is also confirmed in Figure 3.15 (b) and (c) where %*err* and |*err*| are, respectively, shown. The %*err* is presented for both the top and the bottom die at the beginning of the process. This is the moment, during the whole simulated time, in which it reaches its maximum value. This is the coldest

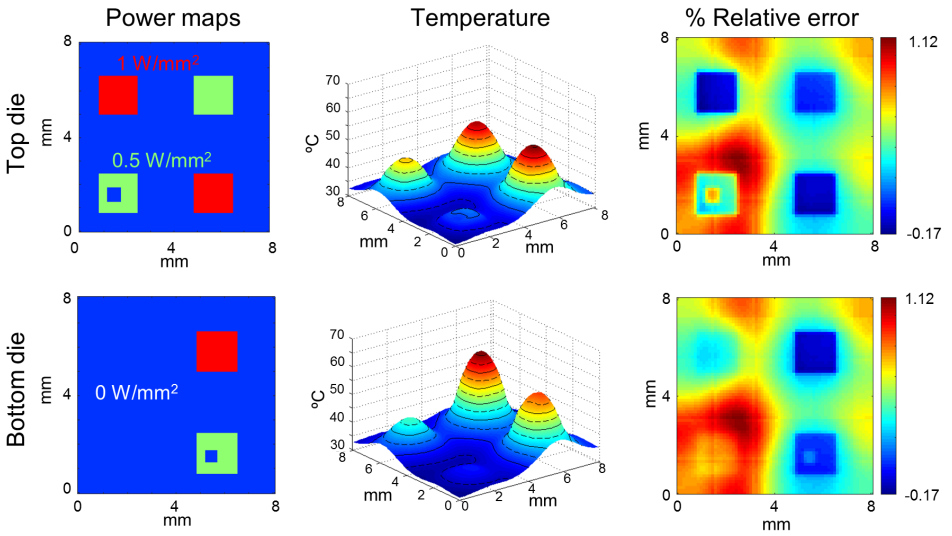


Figure 3.13: Power maps, temperature profiles obtained by FTM and percentage relative error with respect to FEM, for a two dies, face-to-face stack in the steady state regime.

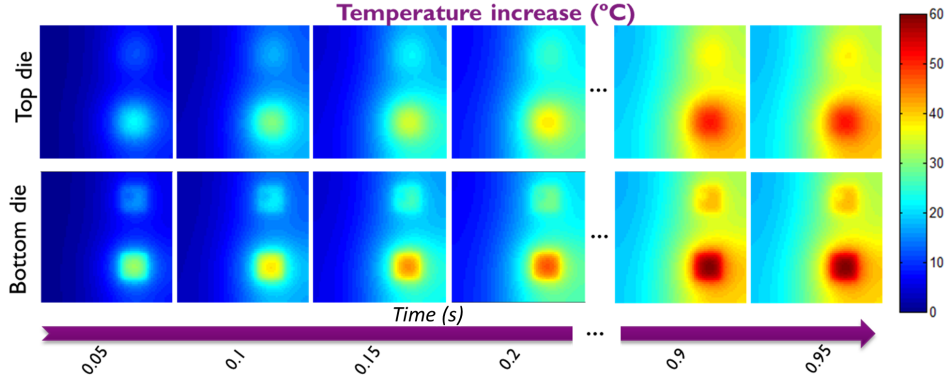


Figure 3.14: Time evolution of the temperature increase profiles on the top and bottom die for constant power dissipation.

stage and the highest values occur in the coldest region of this stage, away from the power dissipation position. The maximum of $|err|$ is, instead, obtained at the end of the simulation, when steady state is almost reached and the highest temperature is experienced, in correspondence with the power dissipation locations. This is shown in Figure 3.15 (c) for both the top and the bottom die. It is worthy to be noted that, in any case, the error with respect to FEM simulations is very low: max($|err|$) is less than 0.17 °C out of approximately 60 °C temperature increase

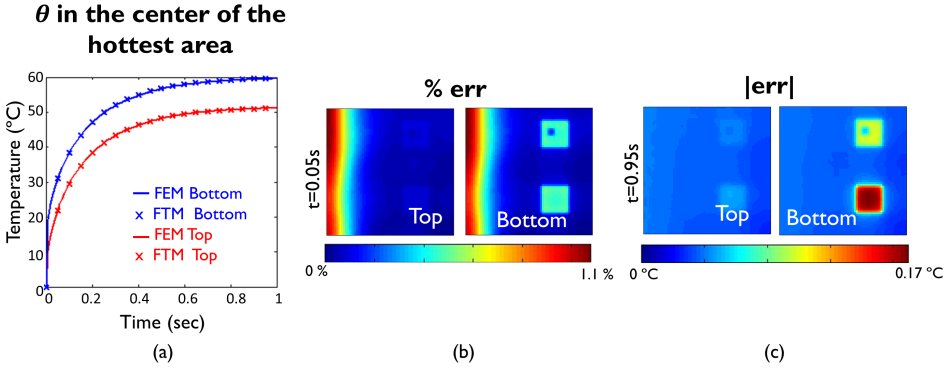


Figure 3.15: (a): Time evolution of the temperature increase in the hottest location of the two dies for the continuous power dissipation scenario in Figure 3.14; comparison between FEM and FTM results. (b): *%err* at the beginning of the process. (c) |*err*| at the end of the simulate period.

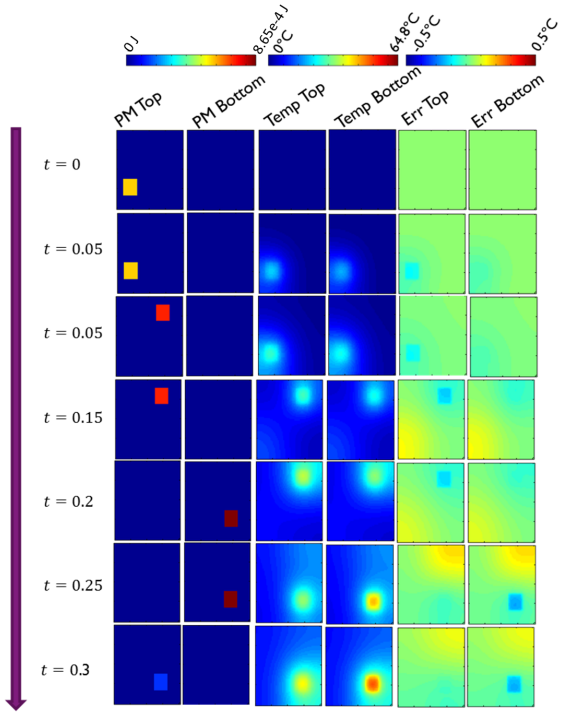


Figure 3.16: The first seven time steps in the evolution of the PM, the temperature increase and the error on the top and bottom die for a case with time varying PMs.

while max(%*err*) is around 1.1%.

Time varying power map

The thermal behavior of the system when subjected to time varying power maps is now analyzed. This case represents a more realistic situation, and, in particular, a kind of situation the transient compact thermal model has been specifically developed for. The simulated time in this case is $t_f = 1.5\ sec$ and $\Delta t = 50\ msec$.

The first seven time steps of the thermal evolution of the system are illustrated in Figure 3.16. The first two columns show the applied PMs on the top and bottom die, the third and fourth columns the temperature increase on the top and bottom die while the last two columns the error, computed as $FEM - FTM$, on the top and bottom die. The absolute value is not considered in this formula in order to understand if the FTM overestimates or underestimates the temperature. From the error plots, the FTM appears to evolve slightly faster than the corresponding FEM: blue-ish error appears during heating up ($FTM > FEM$) while red-ish error during cooling down ($FTM < FEM$). The accuracy is really good, being the error always between $\pm 0.5\ ^\circ C$.

From the Figure, it is clear that the effect of power dissipation is visible on the temperature maps (second and third column in the Figure) starting from one time step after the moment in which power started to be dissipated (first and second column in the Figure). If a certain PM is dissipated in the time interval $[t_1, t_1 + n\Delta t)$, where $n \in \mathbb{N}$, then it is visible in the PM plot in Figure 3.16 referring to the time steps $t_1, t_1 + \Delta t, \ldots, t_1 + (n-1)\Delta t$. The temperature response of the system to that PM, however, starts at $t_1 + \varepsilon$, with $\varepsilon < x,\ \forall x > 0$ (cf. Figure 2.9). This means that, since discrete time is considered, this response can be seen just starting from the plots referring to the time step $t_1 + \Delta t$. It is important to stress that the results of the FTM refer to the point in time that is indicated by the model, without any delay. The delay of the temperature response due to the material capacitance is already included in the HSRs and, therefore, in the FTM results.

Together with accuracy, the other important characteristic of a FTM is its computational speed. As mentioned in Section 2.5.3, the improvement in computational speed between the FTM and FEM approach is case dependent. In case of FEM simulations, indeed, the computational time depends on the complexity, both in time and space, of the PMs and on the mesh size while, for the FTM the computational time is affected by the number of time steps needed to cover the whole simulated time, the spatial resolution and NI. In this particular analyzed case, the computational time for the FTM is almost 300 times faster than analogous FEM (4964 sec FEM with adaptive time step vs 16.7 sec FTM with a fixed time step of 50 msec) with an inaccuracy smaller than $\pm 0.5\ ^\circ C$.

Figure 3.17 summarizes the results of this simulation. It shows data concerning

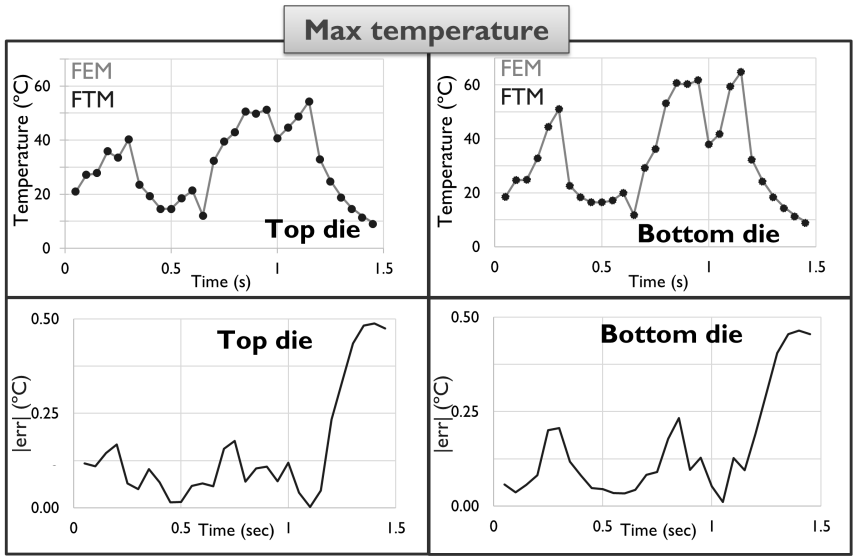


Figure 3.17: Top: maximum temperature on top (left) and bottom (right) die computed by FEM (full line) and FTM (circles) for the case in Figure 3.16. Bottom: $|err|$ on top (left) and bottom (right) die between the maximum temperatures computed by FEM and by FTM.

the maximum temperature achieved in the two dies as a function of time. The left column refers to the top die, while the right column to the bottom die. On the first row the comparison between the temperature increment computed by FEM (line) and by the FTM (circles) is shown. The second row presents two graphs concerning the $|err|$, which is always less than 0.5 °C, in the maximum temperature locations. This shows, once more, that the accuracy of the model is really good. Moreover, the maximum error is located at the end of the process, after 1.2 sec. This is due to the combination of two different reasons. First of all, the steady state of the HSRs is assumed to be reached after 1 sec: this means that the power dissipated more than 1 second before time t_k doesn't have any impact on the temperature increase at time t_k. Moreover, in this simulation, no power is dissipated after 1.1 sec, meaning that, in this set-up, the power dissipated in the past has a higher impact on the temperature increases for $t > 1.1$ sec than in case the chip activity would have been continued. While the chip is active, which is the critical time, the error is less than 0.25 °C.

3.9 Summary

In this Chapter, the *method of images* has been presented as a way to model the thermal behavior of finite dimensional structures, the *die stacks*, starting from normalized temperature responses to hot spot power dissipation in corresponding infinitely large structures. This method is mathematically valid only if an infinite number of images is considered. However, since the value of the HSRs decreases with the distance from the hot spot center, a finite number of them is enough to model with a reasonable accuracy the insulating boundary conditions applied to the lateral sides of the stack. An algorithm to define how many images are needed to model these boundary conditions with a user defined accuracy level has also been presented. Moreover, accuracy assessments on the resolution of the temperature profiles and on the temporal length of the transient HSR have been presented. This last estimation, in particular, is necessary because accounting for the data concerning the evolution of the system until steady state in the HSR can result in a large increase in computational time that is not followed by a corresponding increase in accuracy.

In the last Section of this Chapter, this methodology has been applied to steady state and transient simulations of 3D die stacks with non uniform power maps. It is shown that the FTM is able to predict both the steady state and transient temperature distributions. In both cases, indeed, the results show high accuracy and high reduction in computational time with respect to analogous FEM. More precisely, for the analyzed cases, the percentage error for the steady state regime was less that 1.12% with a 220x speed up in computational time, while, the absolute error in the transient simulation was less than 0.5°C with a 300x speed up. The major steps required in these algorithms are summarized in the flowcharts in Section 3.7.

Part II

Overcoming Limitations of the Stack Model

Chapter 4

Steady State Thermal Impact of μBump Arrays

4.1 Introduction

The limitation of the FTM to model structures constituted by stacked layers of *homogeneous* materials is also related to the requirement of position independent HSRs. The presence of *heterogeneous* material layers is, however, a quite common situation in 3D-technology. The connections between stacked dies are, for example, performed by means of metallic interconnects surrounded by underfill material, while, to allow the connection between PCB and the top dies, copper TSVs are etched in the underlying silicon layers (cf. Section 1.2.2). This means that there exist horizontal layers in which TSVs and silicon and layers in which μbumps and underfill are present at the same time. We focus on the heterogeneity in the interface layers between the chips (μbumps - underfill) and not on the one in the bottom dies (TSVs - silicon) because the ratio between thermal conductivities of the metallic interconnects ($k \approx 30-300\ W/mK$ depending on the amount of Cu, Cu_6Sn_5, Cu_3Sn and Sn after bonding [18, 73]) and the underfill ($k \approx 0.2 - 0.4\ W/mK$ [72] for unfilled or silica filled underfill, up to $k \approx 9 W/mK$ [13] for percolating and neck-based thermal underfill) is much higher than the one between TSVs (Cu, $k = 400\ W/mK$) and silicon ($k = 120\ W/mK$ at 65 °C). As a consequence, the impact of the heterogeneity in the interface layer is much higher than the one in the dies. The thermal impact of the material heterogeneity in the die, within silicon and TSVs, although expected to be small, may be handled in a similar way.

In this Chapter, a methodology is presented to include the steady state thermal impact of specific μbump layouts between the dies. It is applied to face-to-face (F2F), 8 *mm* x 8 *mm* two dies stacks. A F2F configuration is such that the active

regions (faces) of the chips are directed towards each other. In case of a two dies stack, in particular, it means that the active regions are located on the bottom of the top die and on the top of the bottom die. Even if the methodology is presented for this F2F configuration, it could be extended to die stacks of different dimensions as well as to face-to-back (F2B) and back-to-face (B2F) configurations, in which the backside of one chip and the active region of the other chip are facing each other. The added value of the presented algorithm is the inclusion of the local and the global thermal impact of *specific* μbump layouts in the convolution based FTM.

In 3D-IC technology, the μbumps are normally organized in array patterns, with a specific pitch between them. In order to simplify the model, the μbumps are not considered individually in the FTM but equivalent μbump array material properties, which take into account the diameter of the μbumps and the pitch between two of them, are considered. These equivalent material properties are computed by FEM based on a representative volume of the array. They are, therefore, orthotropic and they also include the thermal impact of the underfill that fills the regions between the μbumps in the array (cf. Section 4.2.2 and Figure 4.3). The algorithm discussed in this Chapter has also been published in [58,59] and its basic structure is reported in the flowchart in Figure 4.1. It basically consists in a particular combination of the temperature profiles obtained by assuming homogeneous μbumps and homogeneous underfill material between the two dies. This combination accounts for the specific μbump layout, material properties, geometry and boundary conditions of each considered case.

4.2 Superposition and convolution in case of heterogeneous material layers

The main thermal effect, due to the presence of multiple materials with different thermal conductivities in a single horizontal layer, is the *deviation* of the heat flow lines towards the regions of higher thermal conductivity. This means that the heat path depends on the relative position between the power dissipation locations and the material discontinuities (even without considering the boundary effects). This phenomenon is illustrated in Figure 4.2. The plots show the heat flow lines, through a layer of material, generated by power dissipation in the regions indicated by the thick red lines on top of the models. Convection is assumed on the bottom boundary while insulation on the top one. The modeled geometry is mainly constituted by a material with low thermal conductivity, colored in light blue, with an embedded small rectangular area with higher thermal conductivity, colored in yellow. The former material may represent underfill ($k = 0.4\ W/mK$) while the latter one a μbump array, ($k_x = 0.6\ W/mK, k_y = 4.2\ W/mK$). Note that just a small part of the FEM model is shown in Figure 4.2. Power is dissipated in two different locations: on top of the area corresponding to the material with higher thermal conductivity (labeled "central") in plots (b) and (e), and just next to it

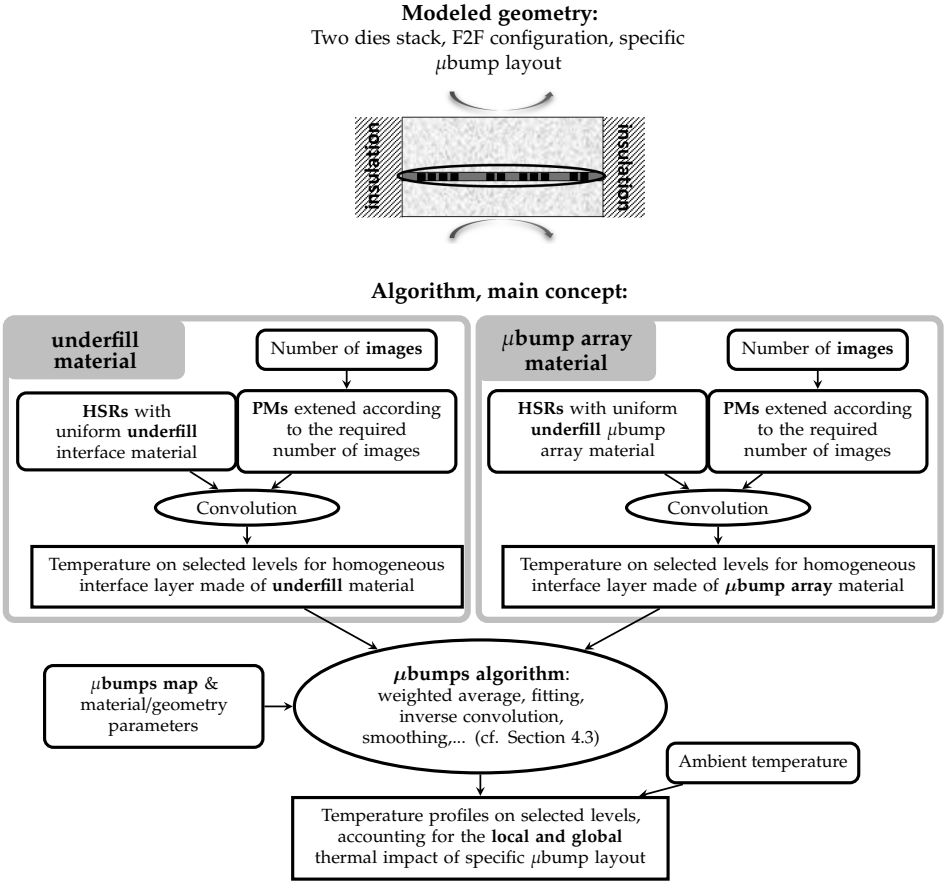


Figure 4.1: Modeled geometry and main concept of the algorithm described in this Chapter.

(labeled "lateral") in plots (a) and (d). Plots (c) and (f) show the situation in which power is simultaneously dissipated in both locations. The idea is to compare the heat flow lines in presence of material heterogeneity and to check the applicability of the superposition principle and of the convolution approach.

Plots (a) and (b) illustrate the effect of material heterogeneity. The black lines are the heat flow lines obtained considering the real structure (different materials) while the magenta, dash lines are computed assuming a homogeneous material layer. More precisely, they are a copy of the heat flow lines in plots (d) and (e). These magenta lines are reported in plots (a) and (b) to highlight the difference between the homogeneous and the heterogeneous material cases. In plot (d) and (e), indeed, just one material is considered for the whole layer: the one below the power dissipation location. This means that the magenta lines in plots (a) and

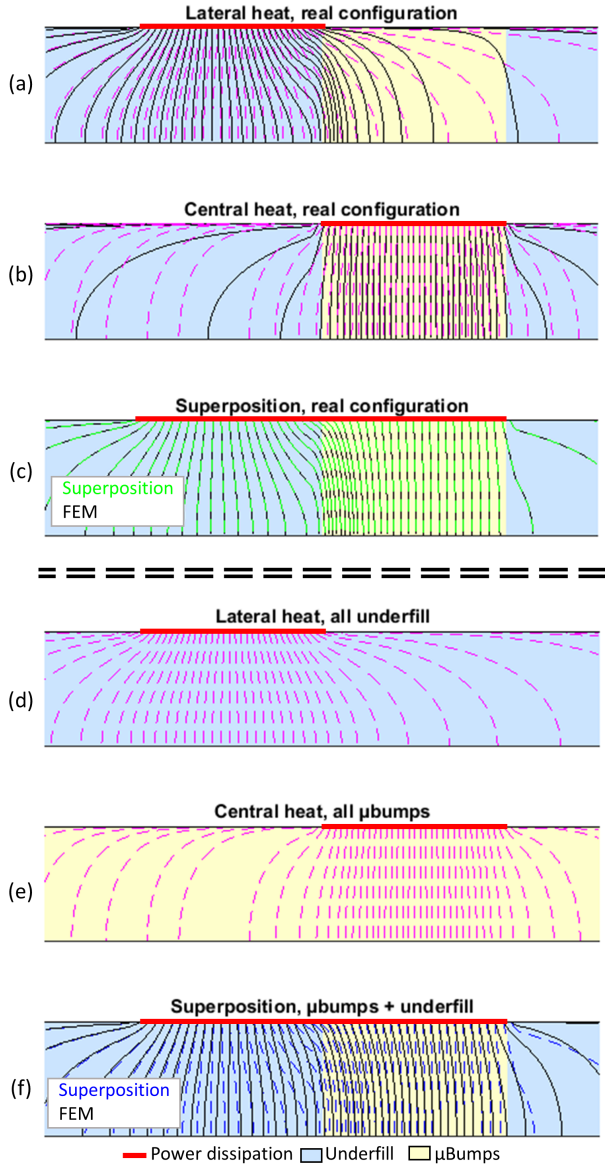


Figure 4.2: Impact of different materials on the heat flow lines. Plots (a), (b) and (c) show the deviation of the heat flow lines towards the region of higher thermal conductivity (yellow) and the applicability of the superposition principle if the individual responses are calculated considering the *real*, heterogeneous structure. Plots (d), (e) and (f) show the responses obtained considering a homogeneous material, the one below the power dissipation, and the error that arises by superposing them in case of material heterogeneity.

(d) are obtained assuming underfill material everywhere, while the ones in plots (b) and (e) assuming equivalent μbump array material everywhere. Even if more materials are present on the same layer (plots (a) and (b)), the steady state heat conduction equation remains linear and, as a consequence, if the heat flow lines obtained for the real structure and separated power sources are superposed, the obtained heat flow lines (green color) correspond to the ones of the FEM simulation (black color).

The issue for the developed FTM, however, arises because convolution is used as the method to apply superposition. As a consequence, the HSRs on each horizontal layer need to be independent of the power dissipation position and this can happen only if the structure is made by layers of homogeneous materials. Plots (d) and (e) show the heat flow lines obtained in case a uniform material (underfill in plot (d) and μbump array in plot (e)) is assumed all over the structure. This is what can be computed by convolving the HSRs, obtained for homogeneous layers, and the dissipated power. When these individual responses are superposed, the resulting heat flow lines (blue color) differ from the ones obtained by FEM simulations (black color). This is because the *deviation* of the heat flow lines induced by the specific heterogeneity is not taken into account by the two homogeneous material models. This example demonstrates that the superposition of the temperature results obtained by assuming homogeneous material layers, according to the location where power is dissipated, doesn't solve the heat conduction problem correctly. This is, however, the result that can be obtained by applying convolution.

It is important to note that this deviation is a *global* effect even if, far enough from the material heterogeneity, it becomes less evident. The distance from the heterogeneity point at which the heat flux deviation becomes negligible is not fixed but it mainly depends on the difference in thermal conductivity between the two materials, the dimensions of the μbump arrays, the BCs and the geometry parameters. Moreover, the thermal effect of the placement of μbumps depends on the sizes and positions of *all* the μbump arrays in the structure: this is why a global approach has to be considered.

Figure 4.2 also clarifies why homogenization techniques [80] are not always useful in this context. Although these strategies provide a good estimation of the *global* thermal impact of having different materials in a single layer, they don't provide any information about the *local* deviation of the heat flow and, therefore, of the temperature profile due to specific heterogeneity layouts. They are mostly used in cases where the relative distribution of the two (or more) materials are periodic and/or uniform. In these cases it is, indeed, reasonable to assign a single equivalent material to the whole layer. Since in microelectronics packages the μbump arrays can be non-uniformly distributed, the goal of the methodology presented in this Chapter is to include both the *global* and the *local* thermal impact of material heterogeneity.

Table 4.1: HSRs generated for the FTM of a two dies stack including the μbumps thermal impact.

HS dissipation\HSR	Top	Bottom
Top	underfill	underfill
	μbump array	μbump array
Bottom	underfill	underfill
	μbump array	μbump array

4.2.1 HSRs generation

Although the superposition principle is still valid, its applicability is highly undermined. Its application would need, indeed, the calculation of HSRs for all the possible combinations of material heterogeneity, i.e. of *μbump layouts*. If the μbump layout is fixed, and a $N_r \times N_c$ discretization grid is used, then $N_r \times N_c$ HSRs are required for each combination [active layer - temperature response layer] to take into account the relative position of each possible HS dissipation location and μbump layout. In case the μbump layout is variable, all the possible combinations [μbumps - underfill] have to be considered. This means the calculation of $2^{N_r \times N_c}$ HSRs. In this way the advantage of being able to model the thermal behavior of the system starting from few 2D-axisymmetric FEM simulations is lost and the pre-processing computational time highly increases. This is the reason why the direct application of the superposition principle has not been considered and a methodology based on convolution has been maintained.

As a consequence, the HSRs are generated considering just uniform material properties in between the two dies. However, as already shown in the previous paragraph, this approach doesn't directly solve the heat conduction problem correctly. The methodology developed for this FTM is based on a specific combination of the HSRs generated assuming uniform material layers whose properties correspond to the individual materials that appear in the heterogeneous layer. In case of underfill and μbump array and for two dies stacks, eight HSRs are generated (Table 4.1). For example, for a hot spot placed on the top die, the HSR is recorded both on the top and the bottom die and, in both cases, one HSR is generated assuming just underfill material and one supposing a full array of μbumps (equivalent properties) in between the two dies. The same for heating up the bottom die.

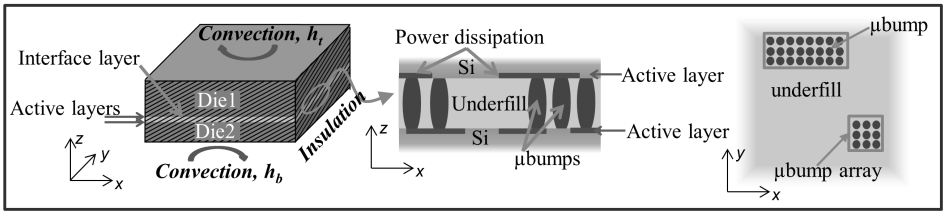


Figure 4.3: Stacked dies and μbump arrays structure. Left: modeled structure. Center: detail of the die-die interface layer. Right: top view of μbump arrays embedded in underfill.

4.2.2 Modeling of interface material layer

The modeled structure, for which the steady state extension of the FTM to include the μbumps thermal impact has been developed, is illustrated in Figure 4.3. On the left hand side, the two dies stack is presented, the positions of the active layers (face-to-face configuration), which are also the temperature computation layers, are highlighted. The applied boundary conditions are also indicated: equivalent convection on the top and bottom sides and lateral insulation. In the central part of the picture, the vertical cross section, zoomed-in around the interface layer, is shown. Finally, on the right hand side, the top view of the interface layer is presented.

The μbumps are normally organized in array patterns. Examples of possible patterns for the μbump arrays are the *area array*, where the μbumps are uniformly distributed over the whole chip area and the *peripheral array*, where the μbumps create a frame close to the edges of the stack. In other situations, the placement of the μbumps can be associated with the location of certain functional blocks in the chip. In all these cases, since the diameter of the single μbump and the pitch between two of them are very small (diameter $7.5 - 25\mu m$, pitch $20 - 50\mu m$ [29,49]) with respect to the die size (*mm* or *cm*) and since the placement of the μbumps *within* the array follows a regular pattern, homogenization techniques can be used to obtain *equivalent μbump array* material properties. This means that the interconnects are not included individually but equivalent material properties are assigned to each array, taking into account individual μbumps dimensions and their in-between distance or pitch. Both the equivalent in-plane and out-of-plane thermal conductivities are computed by means of FEM. They are obtained by matching the thermal behavior of a heterogeneous unit cell, which is repeated periodically in the array, with the one of a cell constituted by a single homogeneous material [31,73]. This does not significantly affect the temperature results [118] while avoiding the use of an extremely fine grid size, which would cause higher computational time.

It should be noted that, due to the structure of the μbump arrays, the equivalent

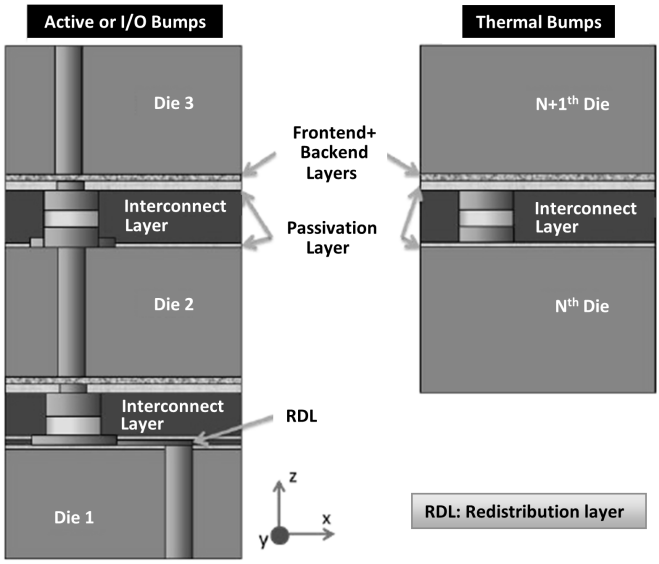


Figure 4.4: Architecture and key interfaces associated with active and dummy (thermal) μbumps (from [18]).

μbump array material properties are orthotropic. More precisely, the in-plane thermal conductivity of the μbumps is lower than the out-of-plane one. This is because the metallic interconnects have a preferred vertical direction: each of them, indeed, vertically connects the two silicon dies and is surrounded by underfill material with a low thermal conductivity. To simplify the notation, in the following of this Chapter, the term *μbumps* refers to *μbump arrays* and the corresponding material properties are the equivalent ones computed for the μbump arrays. These equivalent properties include, therefore, the impact of the underfill material that fills the regions between the μbumps in the array.

Furthermore, there are two different categories of μbumps: *functional* (or *active*) μbumps, which electrically connect two dies, and *dummy* (or *thermal*) μbumps, which merely improve the thermal performances of the device. As reported in Figure 4.4, these two kinds of interconnections are structurally different and, as a consequence, they perform differently from a thermal point of view [18]. To allow for electrical connections, indeed, the functional μbumps are connected to metals (TSVs, redistribution layers,. . .) with high thermal conductivity, while the dummy μbumps are not. In this Chapter, however, this difference is not taken into account and it is assumed that the dummy μbumps and the functional μbumps thermally behave in the same way.

Table 4.2: System parameters and their ranges for which the FTM to include the μbumps thermal impact has been developed.

Parameter	Min possible value	Max possible value
h_t, heat transfer coeff. top (W/m^2K)	0	15000
h_b, heat transfer coeff. bott. (W/m^2K)	0	15000
$h_t + h_b$ (W/m^2K)	500	15000
k_z μbumps, out-of-plane equivalent μbumps thermal conductivity (W/mK)	2.5	20
$k_{x,y}$ μbumps, in-plane equivalent μbumps thermal conductivity (W/mK)	0.3	13.5
k underfill, thermal conductivity underfill (W/mK)	0.2	9
l top die, thickness top die (μm)	20	300
l bottom die, thickness bott. die (μm)	20	300
l interface, thickness interface (μm)	2	30
$\tilde{\rho}$, area μbump array/total area	0.09	0.5

4.2.3 Degrees of freedom

In this frame, the values of the system parameters, such as the top and bottom heat transfer coefficients, the material properties and the thicknesses of the layers, can be freely selected by the user within certain ranges reported in Table 4.2, allowing the analysis of their thermal impact. The ranges in Table 4.2 cover most of the common microelectronics applications set-ups [13,73]. The fitted model presented in this thesis is valid for values of the system parameters within these ranges; if some other configurations outside this framework need to be analyzed, the whole fitting procedure has to be re-run including the new values of the parameters.

As long as the area ratio, $\tilde{\rho}$, between the μbump arrays and the underfill is in between the values presented in Table 4.2, the location and the layout of the different arrays can be freely chosen. This is of particular importance since, having the interconnections a much higher thermal conductivity than the underfill material, additional dummy μbumps can be inserted merely on a thermal basis to improve heat dissipation, without providing any electrical connection. However, due to the high additional costs, a thermally optimized selection of the μbumps

number and placement is essential. This means that the model should be able to deal with different μbump amounts and layouts. In this frame, it is important to note that this methodology has been implemented considering a maximum of two equivalent material properties (referred to as μbumps and underfill) for the interface layer. This means that just one combination of μbumps diameters and pitches is considered. No limitations, instead, are imposed on the dissipated power location and intensity.

4.3 FTM methodology to include the thermal impact of μbump arrays

As shown in Section 4.2, the superposition principle can be applied to include the thermal impact of local material heterogeneity. However, in order to keep the freedom to select the power maps and the μbumps distributions, the computation of dedicated HSRs for each specific μbump layout and HS power dissipation position is required. The advantage in reducing the computational cost and in the possibility to obtain the temperature response of the system by means of couple of simple FEM simulations is, therefore, lost. This is the reason why the developed FTM includes the thermal impact of the μbump arrays *after* convolution, by properly combining the two temperature results obtained by assuming uniform underfill and uniform μbumps interface materials.

The development of this methodology has been carried out starting from a simplified situation in which uniform power is dissipated on the top die and insulation is assumed on the top of the stack. Extensions to more general situations are progressively introduced. More precisely, the generalization steps are the following:

1. Uniform power dissipation on the top die, convection allowed just from the bottom side of the stack. Since all the heat is removed from the bottom side of the stack, it has to go through the interface layer with heterogeneous material properties;
2. Uniform power dissipation on the top die, convection allowed both from the *top* and the bottom side of the stack;
3. Uniform power dissipation on both the top and *bottom* die, convection allowed both from the top and the bottom side of the stack;
4. *Non-uniform* power dissipation on both the top and bottom die, convection allowed both from the top and the bottom side of the stack.

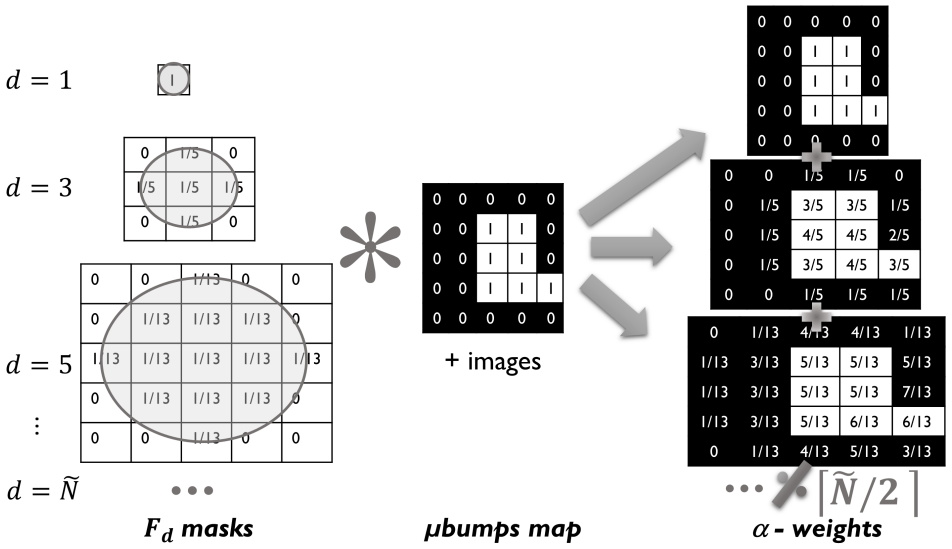


Figure 4.5: α-weights computation technique. The α-weights account for the reciprocal position of the μbump arrays. They are used to compute the temperature profile as the weighted average between $\Theta_{11,und}$ and $\Theta_{11,\mu b}$.

4.3.1 Uniform power on top die, convection from bottom side

Temperature on top die, weighted average

In a first approximation, the temperature increase profile on the top die, $\tilde{\Theta}_{11}$ in equation (4.3), is calculated through a *weighted average* between $\Theta_{11,und}$ and $\Theta_{11,\mu b}$. These temperature profiles are obtained by convolution using, respectively, the HSRs (both temperature response and HS dissipation on the top die) for uniform underfill (*und*) material and uniform μbumps (μb) material. In the following, the digits in the subscripts refer, respectively, to the HS dissipation layer and to the temperature computation layer; *1* indicates the top die and *2* the bottom die. In case of only one index, it refers to the temperature computation layer taking into account the dissipated power on both dies.

The weights, $\alpha(i, j)$, depend on the overall μbump layout and are grouped in a $N_r \times N_r$ matrix ($N_r = N_c$ since square dies are considered), which is of the same size as the ones storing the temperature responses data and the *μbumps map*. This last one, in particular, is a binary matrix indicating the die-die interface material considered in each cell: 0 indicates underfill and 1 a part of a μbump array. More precisely, the values in the α-matrix are computed averaging the results of numerous convolutions between masks of increasing size, but of unitary sum, and

the μbumps map (cf. Figure 4.5). Each mask F_d is defined as:

$$\begin{aligned} &F_d \in Mat(d \times d), \quad \text{with } d = 1,\ 3,\ 5, \dots \\ &F_{d,ij} = \varphi, \quad \text{if } \sqrt{(i - \lceil d/2 \rceil)^2 + (j - \lceil d/2 \rceil)^2} \leq \lfloor d/2 \rfloor \\ &F_{d,ij} = 0, \quad \text{if } \sqrt{(i - \lceil d/2 \rceil)^2 + (j - \lceil d/2 \rceil)^2} > \lfloor d/2 \rfloor \\ &\sum_i \sum_j F_{d,ij} = 1 \end{aligned} \tag{4.1}$$

where i, j are row and column indexes ($-\lfloor d/2 \rfloor \leq i, j \leq \lfloor d/2 \rfloor$) and φ is a constant value that is determined by solving the last equation above. $\lceil x \rceil$ indicates the smallest integer number greater than or equal to x while $\lfloor x \rfloor$ is the greatest integer number smaller than or equal to x. In other words, the value of the entries in each F_d matrix is greater than zero just inside a circle of radius $\lfloor d/2 \rfloor$ and the sum of all its entries is 1. The matrix of weights, α, is, then, computed as the central $N_r \times N_r$ portion of

$$\alpha = \frac{\sum_{k=1}^{\lceil \tilde{N}/2 \rceil} F_d *_{2D} \mu\text{bumpsMap}^e}{\lceil \tilde{N}/2 \rceil}, \quad d = 2k - 1, \quad \tilde{N} = \lceil \sqrt{2\pi} N_r \rceil \tag{4.2}$$

where μbumpsMap^e is the extension of the μbumps map via the method of images (Chapter 3) and N_r indicates the number of rows (and columns) in the temperature response matrices. $\tilde{N}$ is selected so that the area of a circle, whose radius is equal to the diagonal of the die, is equal to $\tilde{N}^2 \bar{h}^2$ ($\bar{h}$ is the grid size). In this way, the largest F_d mask is big enough to ensure that each element in the α-matrix is influenced by the overall μbump arrays layout plus some edge effect. This approach allows to include the global impact of each specific μbump layout by weighting the material impact by the relative distances between cells.

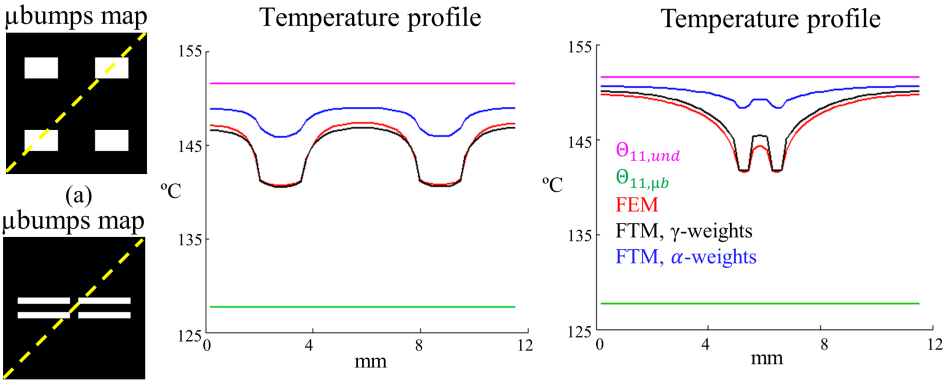
The temperature profile is, then, computed as

$$\tilde{\Theta}_{11} = \alpha \Theta_{11,\mu b} + (1 - \alpha) \Theta_{11,und}. \tag{4.3}$$

The results for the μbumps maps (black=underfill, white=μbump array) shown in Figure 4.6 (a) and (b) are presented in blue in Figure 4.6 (c) and (d) respectively. They refer to the diagonal cross sections indicated by the yellow dashed lines in plots (a) and (b). The red curves in the same graphs represent the FEM results, against which the model has been validated. The values of the design parameters used in these examples are also reported in the Figure.

Temperature on top die, fitting procedure

From Figure 4.6 it is clear that this methodology is not able to satisfactory describe the phenomenon because of the complicated relationship, which acts *deviating* the



Parameters:

h_b	$k_{\mu b,xy}$	$k_{\mu b,z}$	k_{und}	l_t	l_b	l_i	q
8000 (W/m^2K)	0.75 (W/mK)	6 (W/mK)	0.5 (W/mK)	200 (μm)	50 (μm)	13 (μm)	1 (W/mm^2)

Figure 4.6: (c), (d): Temperature profiles on the diagonal of the top die, in a 8×8 mm^2 two dies stack, for uniform power dissipation on the top die and two different μbump layouts ((a), (b) respectively). Comparison between the α-weights FTM (blue), the γ-weights FTM (black) and the FEM (red) results. $\Theta_{11,und}$ (magenta) and $\Theta_{11,\mu b}$ (green) are also indicated.

heat flow, between the temperature drop due to μbumps, all the system parameters and the μbump layout itself. To overcome this issue a fitting approach based on multiple FEM solutions has been performed.

As a first step, the temperature profiles on the top die, obtained via FEM models by dissipating uniform power on the top die, have been recorded for more than 130 different sets of parameters and 60 μbump layouts. The values of the design parameters, which are listed in Table 4.2, have been chosen (uniformly distributed) between their allowed maximum and minimum value. Then, for each temperature result, new $\tilde{\gamma}$-weights are computed so that

$$\Theta_{FEM,ij} = \tilde{\gamma}\Theta_{11,und} + (1 - \tilde{\gamma})\Theta_{11,\mu b} \tag{4.4}$$

where $\Theta_{FEM,ij}$ is the temperature increase at position i, j obtained by FEM. $\tilde{\gamma}$ are, in fact, the *exact* weights needed in the average and, for a fixed μbump layout and system geometry, they are position dependent. More precisely, they are directly computed from the temperature increase values as

$$\tilde{\gamma} = \frac{\Theta_{FEM,ij} - \Theta_{11,\mu b}}{\Theta_{11,und} - \Theta_{11,\mu b}}. \tag{4.5}$$

At a second stage, the $\tilde{\gamma}$ values obtained from each specific FEM simulation are plotted with respect to the corresponding α-weights. In Figure 4.7, these results are

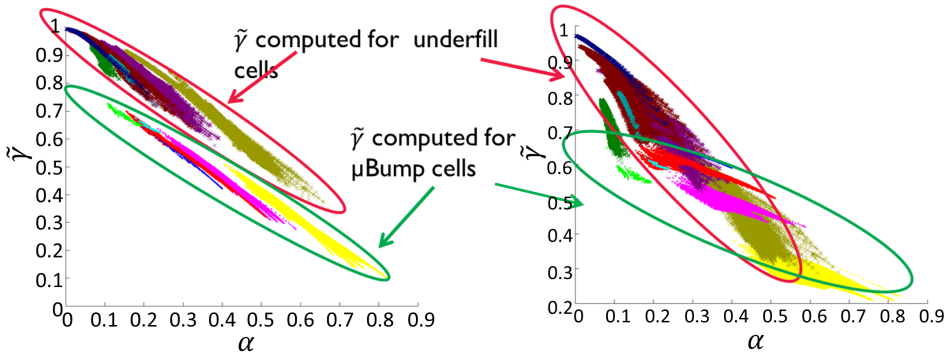


Figure 4.7: $\tilde{\gamma}$ values vs α values; different colors indicate different μbump array area ratios $\tilde{\rho}$; darker tones refer to underfill cells while lighter tones to μbumps cells.

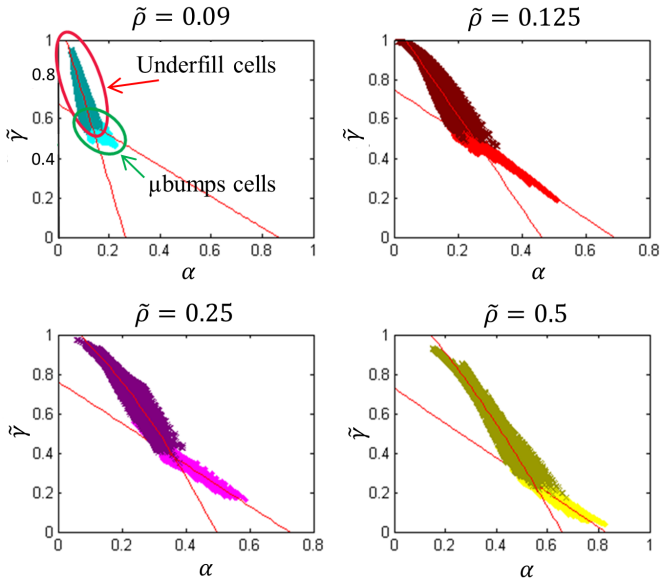


Figure 4.8: Fitting of the $\tilde{\gamma}$-weights with respect to the α-weights for one particular set of system parameters and four different $\tilde{\rho}$ values. Darker colors refer to underfill cells while lighter colors to μbump cells.

shown for two particular cases. As a first comment, we can note from the Figure that $\tilde{\gamma} < (1 - \alpha)$, meaning that, as already inferable from Figure 4.6, the α-weights FTM underestimates the μbumps thermal impact or, from another point of view, it overestimates the temperature increase. The approach presented in this Section uses this information to improve the results.

In each of the two graphs in Figure 4.7, all the system parameters are kept constant except for the μbumps area ratio, $\tilde{\rho}$. The clustering according to the $\tilde{\rho}$ value is shown by using different colors. It is important to note that, for the same $\tilde{\rho}$ value, different μbump layouts are considered. Moreover, for the same color, a darker tone indicates $\tilde{\gamma}$ values referring to *underfill* cells in the μbumps map while a lighter tone to μbumps cells. From the plots it is, indeed, clear that the weights behave differently depending on the nature (μbump array or underfill) of the cell they refer to. As a consequence, the fitting of the $\tilde{\gamma}$-weights with respect to the α-weights is performed distinguishing between the nature of each single cell. More precisely, quadratic polynomial, least square fittings ($\gamma = a\alpha^2 + b\alpha + c$) are used for underfill cells while linear least square fittings ($\gamma = d\alpha + e$) are sufficient for μbumps cells. Examples of these fittings for different $\tilde{\rho}$ values are shown in Figure 4.8: darker colors refer again to underfill cells while lighter colors to μbumps cells.

Finally, the overall quadratic fittings of the a, b, c, d, e parameters with respect to the system parameters provide five different models that allow retrieving the γ values for different simulation setups. The top temperature profile can, then, be computed from

$$\Theta_{11} = \gamma\Theta_{11,und} + (1-\gamma)\Theta_{11,\mu b} \tag{4.6}$$

where γ is calculated using the underfill related model in cells corresponding to underfill material and the μbumps related model elsewhere. The new temperature estimations, which approximate the FEM results much better, are shown in black in Figure 4.6 (c) and (d).

Defining the relative temperature error as

$$err = \frac{\Theta_{FEM} - \Theta_{FTM}}{\Theta_{FEM}}, \tag{4.7}$$

where Θ_{FTM} is the temperature increase obtained by the FTM, the maximum of its absolute value drops from 4% to 0.35% in (c) and from 4.8% to 0.96% in (d) considering the γ-weights instead of the α-weights model. The increase in computational time, once the fitting models have been established, is negligible. At first sight, the original α-weights error could be acceptable. It is, however, important to highlight that this definition of the error is *global*, meaning that it is mainly influenced by the variation in the thermal resistance of the interface layer with respect to the *overall* thermal resistance. Since $R_{th,interface} \ll R_{th,system}$, the global error is small already by using the α-weights approach. However, the error can also be defined in a more *local* way. It can, for instance, be normalized with respect to the difference in the thermal resistance of the interface layer in case of uniform underfill and uniform μbumps material (i.e. with respect to the impact of the interface material). Doing so, the improvement in accuracy achieved by using the γ-weights FTM becomes more clear. More precisely, by defining the local error as

$$err_l = \frac{|\Theta_{FEM} - \Theta_{FTM}|}{\Theta_{11,und} - \Theta_{11,\mu b}}, \tag{4.8}$$

the local error is reduced from 25%, for the α-weights FTM, to less than 5%, for the γ-weights FTM, for the two cases in Figure 4.6.

The main drawbacks associated to the γ-weights methodology are that, since it relies on fitting,

- it is only valid for the ranges of parameters in which the fitting has been performed (Table 4.2);
- the physical meaning of the phenomenon is lost.

However, within the valid parameters ranges, the fitting coefficients are computed only once and they can be used to model all the μbumps patterns.

Temperature on bottom die

The computation of the temperature increase profile on the top of the bottom die, Θ_{12}, for the same simplified situation cannot be handled in the same way. For uniform heating of the top active layer and cooling at the bottom of the die stack, the heat path is one dimensional, vertical and downwards. In this case, the choice of the *uniform* interface material (μbumps or underfill) doesn't have any impact on the obtained temperature profile on the bottom die. The interface layer between the top and bottom die is, indeed, outside the section of the heat flow path between the point in the bottom die where the temperature is evaluated and the end of the heat path (heat extraction at the bottom side of the stack). Any *uniform* material change upstream this temperature evaluation location doesn't have, in this setting, any impact on the temperature. This results in $\Theta_{12,und} = \Theta_{12,\mu b}$. A schematic illustrating this situation, in terms of a resistance network, is shown in Figure 4.9. However, if a particular μbump layout is assumed, the heat path is deviated towards the areas with higher thermal conductivity. This means that it is no longer strictly vertical and a non-uniform temperature profile, which cannot be achieved by any combinations of $\Theta_{12,und}$ and $\Theta_{12,\mu b}$, is experienced.

The computation of Θ_{12} starts, therefore, from Θ_{11}. For a general power map on the top die, PM_1, and uniform die-die material, Θ_{11} and Θ_{12} can be computed, using the basic methodology for the stack configuration, as

$$\Theta_{11} = PM_1 *_{2D} HSR_{11}, \qquad \Theta_{12} = PM_1 *_{2D} HSR_{12}. \tag{4.9}$$

Manipulating equations (4.9) and defining $\beta = HSR_{12} /^{(*)} HSR_{11}$ (where $/^{(*)}$ indicates inverse 2D-convolution), it is possible to compute Θ_{12} starting from Θ_{11} and without any prior knowledge about PM_1:

$$\Theta_{12} = PM_1 *_{2D} HSR_{12} = PM_1 *_{2D} HSR_{11} *_{2D} \beta = \Theta_{11} *_{2D} \beta. \tag{4.10}$$

Although this is useless in case of uniform die-die material since PM_1 is an input parameter and, therefore, it is known, this relationship can be exploited when particular μbump layouts are considered: the effect of heterogeneous die-die material to *deviate* the heat path can, indeed, also be viewed as *power redistribution*.

By exploiting equation (4.10), the two temperature profiles $\Theta_{12,und}$ and $\Theta_{12,\mu b}$ are computed using, respectively, the HSRs obtained for underfill and μbumps material to compute β. More precisely,

$$\Theta_{12,und} = \Theta_{11} * \beta_{und}, \qquad \Theta_{12,\mu b} = \Theta_{11} * \beta_{\mu b} \tag{4.11}$$

where $\beta_{und} = HSR_{12,und}/^{(*)}HSR_{11,und}$ and $\beta_{\mu b} = HSR_{12,\mu b}/^{(*)}HSR_{11,\mu b}$.

The two obtained temperature profiles are, then, combined according to

$$\Theta_{12} = \Theta_{12,und} \cdot (1 - \mu\text{bumpsMap}) + \Theta_{12,\mu b}\mu\text{bumpsMap}. \tag{4.12}$$

This means that the temperature values obtained considering the quantities related to the underfill are used in the cells corresponding to underfill, while the ones obtained considering quantities related to μbumps, in the μbumps cells. This step results in a highly un-smooth profile (blue curves in Figure 4.10 (a) and (b)). This is because the curvature of the top temperature profile is opposite to the one of the bottom and the selected profiles combination methodology is not considering neither this fact neither a transition area.

A final smoothing step is introduced to compensate for this un-smoothness. After several trials, a smoothing approach based on two Gaussian low pass filters has been selected. A narrower Gaussian filter is used to smooth out the regions corresponding to μbumps cells since the heat, that is already over a μbump cell on the top die, is more likely to go through the μbump itself and less influenced by its neighbors. The heat that is over an underfill cell, instead, is more likely to spread and, therefore, a wider filter is used. The two variance parameters are selected through an optimization procedure over 51 different parameters sets and 14 μbump layouts. The values that reduce the Pseudo-Huber error the most have been selected. This metric, which is defined as

$$Huber = 4 \sum \left[\sqrt{1 + \left(\frac{err_{i,j}}{2}\right)^2} - 1 \right] \tag{4.13}$$

with $err_{i,j}$ the relative error in cell (i, j) with respect to FEM results, has been selected in the optimization procedure since, in this penalty function, the outliers are less relevant. The selected values for the variances of the Gaussian filters, obtained by minimizing the Pseudo-Huber loss function, are fixed to 768e-6 for underfill and 84.7e-6 for μbumps. The decision to keep these values fixed is based on the results of a sensitivity analysis. This analysis has shown that the dependence of the best possible variances values, on the different system parameters and μbump layouts, is low enough to be negligible without any significant impact on accuracy.

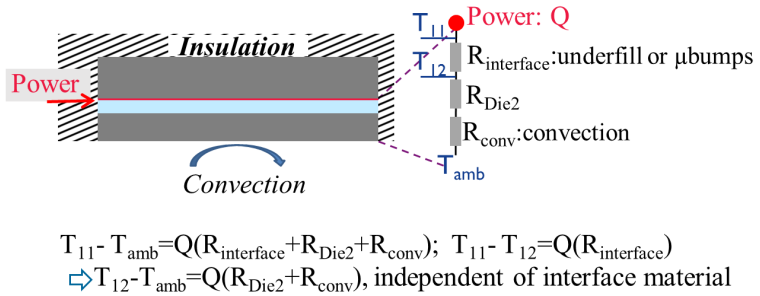


Figure 4.9: Schematic of the reason why $\Theta_{12,und} = \Theta_{12,\mu b}$ for uniform interface material, power dissipation on the top die and heat removal from the bottom side.

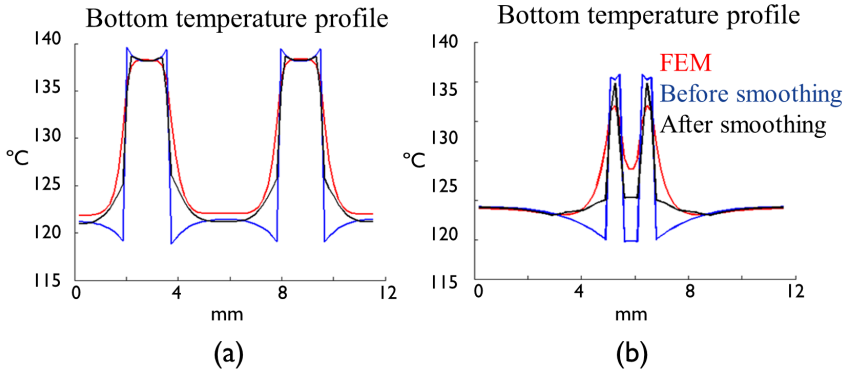


Figure 4.10: Temperature profiles on the diagonal of the bottom die for the μbumps maps shown in Figure 4.6 (a) and (b) respectively; red lines refer to FEM results, blue lines to FTM before smoothing and black lines to FTM after smoothing.

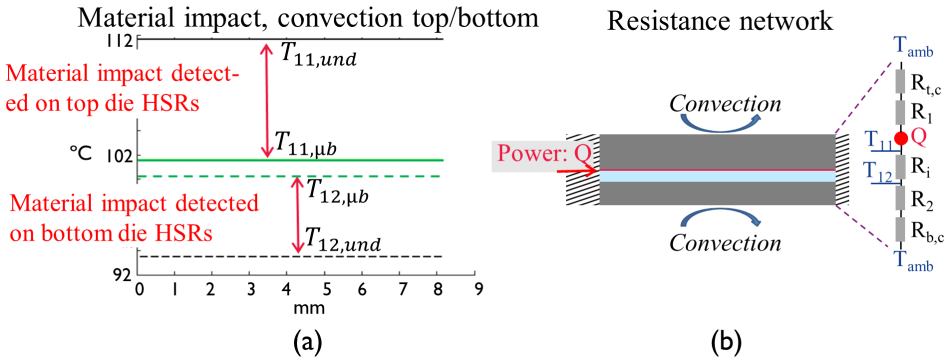


Figure 4.11: (a): Impact of the uniform die-die interface material on the temperature profiles in case of convective boundary condition both on top and bottom. (b): Resistance network used to compute the temperature increase in case of homogeneous interface material.

The considered smoothing masks are, then, normalized to sum up to one, since just a redistribution of the temperature is needed, not a change of its value. The final results are shown in black in Figure 4.10 (a) and (b). The graphs show the clear improvement introduced by the smoothing: the issue of having temperature profiles with opposite curvatures than the corresponding FEM results is solved. Moreover, the overall accuracy with respect to FEM is highly increased.

4.3.2 Uniform power on top die, convection from both sides

In the second step of the methodology development, the situation is considered in which convective boundary conditions are applied both on the top and the bottom of the stack through uniform heat transfer coefficients, h_t and h_b, and uniform power is dissipated on the top die. In this setting, the thermal impact of the *uniform* die-die interface material can be detected both on the top and on the bottom temperature profiles (Figure 4.11 (a)). This is because two vertical heat paths, upwards and downwards, are now present. The thermal conductivity of the uniform interface material has an impact on how the heat flux splits, influencing the uniform temperature both on the top and the bottom die. More precisely, describing the heat conduction phenomenon via the corresponding resistance network, the temperature increases on the top and the bottom dies can be calculated as follows

$$\Theta_{11} = Q\frac{(R_1 + R_{t,c})(R_i + R_2 + R_{b,c})}{R_1 + R_{t,c} + R_i + R_2 + R_{b,c}}, \qquad \Theta_{12} = Q\frac{(R_1 + R_{t,c})(R_2 + R_{b,c})}{R_1 + R_{t,c} + R_i + R_2 + R_{b,c}}, \quad (4.14)$$

where R_1, R_i, R_2 are the conductive thermal resistances of the top die, the interface layer and the bottom die respectively, $R_{t,c}$ and $R_{b,c}$ the convective thermal resistances on top and bottom of the stack and Q the dissipated power (cf. Figure 4.11 (b)). The term R_i, which appears in both equations, depends on the thermal conductivity of the interface layer and, therefore, $\Theta_{11,und} \neq \Theta_{11,\mu b}$ and $\Theta_{12,und} \neq \Theta_{12,\mu b}$. The thermal impacts, Eff_1 and Eff_2, of the uniform interface materials on the top and the bottom die can, therefore, be respectively defined as

$$Eff_1 = \Theta_{11,und} - \Theta_{11,\mu b} \qquad Eff_2 = \Theta_{12,\mu b} - \Theta_{12,und}. \quad (4.15)$$

In the situation discussed in the previous Section where $h_t = 0$, if a particular μbump layout was introduced in the interface layer, the temperature on the top die was always in between $\Theta_{11,und}$ and $\Theta_{11,\mu b}$, i.e. $\Theta_{11,\mu b} \leq \Theta_{11} \leq \Theta_{11,und}$. However, if the heat flux splits in two different paths, it is possible that $\Theta_{11} > \Theta_{11,und}$ and $\Theta_{11} < \Theta_{11,\mu b}$ (Figure 4.12). This phenomenon depends on different factors and it originates from the splitting of the heat flux, from the spreading/constriction resistances due to material heterogeneity in the interface layer and from the presence of the top and bottom dies. The thicknesses of the silicon layers play, indeed, a relevant role in this situation, creating the room for constriction and

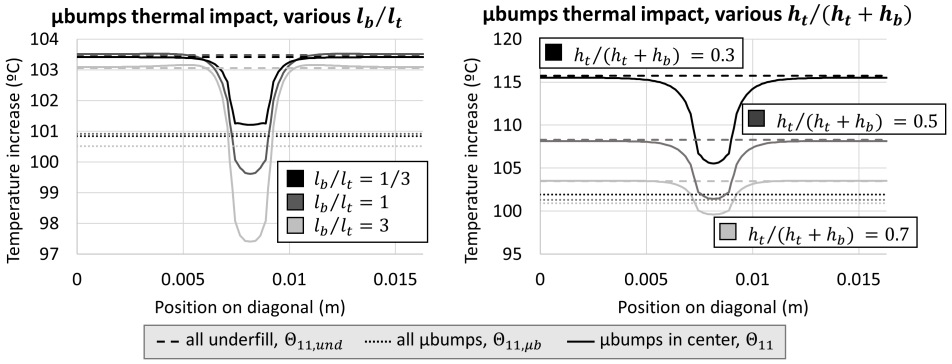


Figure 4.12: Effect on Θ_{11} of a heterogeneous interface layer. Full lines represent cases in which a μbump array is surrounded by underfill, dashed lines are for full homogeneous underfill layer and pointed lines for area arrays. Different colors represent different system parameters and, on the left hand side, the thicknesses of the dies are changed (fixed $h_t = 7000\ W/m^2K$, $h_b = 3000\ W/m^2K$), while on the right hand side, the convection coefficients are varied (fixed $l_b = l_t = 150\ \mu m$).

spreading to occur. A thick top die increases the possibility of the heat flow to be *attracted*, while still in the top die, towards the area with higher interface thermal conductivity (higher constriction resistance). A thick bottom die, on the other hand, creates more room for the heat flow to spread before being convectively transfered to the ambient (higher spreading resistance). This basically means that the ratio between the bottom and the top die thickness, $\frac{l_b}{l_t}$, which gives and indication of the ratio between spreading and constriction resistances, plays an important role in determining the effect of a specific μbumps heterogeneity on Θ_{11}. The spreading in the bottom die is, in particular, responsible for $\Theta_{11} < \Theta_{11,\mu b}$.

This, however, can happen only if the heat flow splits in two different parts. In the situation analyzed in the previous Section where $h_t = 0$, 100% of the heat flux flew towards the bottom and $Eff_2 = 0$. The impacts of the spreading/constriction resistances were already included in the γ coefficients, the bottom heat path was *saturated* since all the heat was going downwards and no extra possibility of temperature reduction was provided by Eff_2. In case the heat path splits in two different sections, the heat flux towards the bottom is not *saturated* and more spreading is allowed. In this situation, another important role, in determining the effect of a specific μbumps heterogeneity on Θ_{11}, is played by the parameter $\frac{h_t}{h_t+h_b}$, which gives an approximation of the fraction of heat that is removed from the top surface of the stack. This is because, in case of uniform interface material, Q_1 and Q_2, which represent, respectively, the amount of heat flowing upwards and downwards, can be computed, using a resistance network approach, as

$$Q_1 = Q\frac{R_i + R_2 + R_{b,c}}{R_1 + R_{t,c} + R_i + R_2 + R_{b,c}}, \qquad Q_2 = Q\frac{R_1 + R_{t,c}}{R_1 + R_{t,c} + R_i + R_2 + R_{b,c}}. \tag{4.16}$$

However, since the internal conductive resistances $R_{th} = \frac{l}{kA}$ normally ranges between $\frac{6.5e-5}{A}$ $°C/W$ and $\frac{1.5e-6}{A}$ $°C/W$, while the convective resistances are normally higher than $\frac{1e-4}{A}$ $°C/W$, the previous quantities can be approximated as

$$Q_1 \approx Q\frac{h_t}{h_t + h_b}, \qquad Q_2 \approx Q\frac{h_b}{h_t + h_b}. \tag{4.17}$$

This means that the ratio $\frac{h_t}{h_t+h_b}$ gives an approximation of how the heat flow splits in case of uniform interface material having, therefore, this ratio and not h_t or h_b separately, a significant impact in determining Θ_{11}.

Considering the value of $\max(\Theta_{11,\mu b} - \Theta_{11})$, it increases, in particular, with the thickness l_b of the bottom die, which provides more room for spreading, and with the relative amount of heat, $\frac{h_t}{h_t+h_b}$, that is removed from the top surface of the stack. This increases the possibility of the spreading to happen since the relative amount of heat that goes downwards decreases. Moreover, it is important to note that the extra spreading and the possibility to have $\Theta_{11} < \Theta_{11,\mu b}$ is also allowed by the fact that, in case $h_t > 0$, $Eff_2 > 0$ or, in other words, the temperature in the underfill locations on the top of the bottom die is expected to be lower than in the μbumps locations (Figure 4.11 (a)).

After this analysis, we can conclude that the impact on Θ_{11} of a particular μbump layout depends on the thermal impacts of uniform interface materials both on the top and the bottom die (Eff_1 and Eff_2), on the ratio between the thicknesses of the dies $\left(\frac{l_b}{l_t}\right)$, on the relative amount of μbumps cells vs underfill cells ($\tilde{\rho}$) and on the fraction of heat that is removed from the top surface of the stack $\left(\frac{h_t}{h_t+h_b}\right)$. Accounting for this last parameter, in the analysis concerning the thermal impact on Θ_{11} of specific μbump layouts in different situations, which will follow in this Section, the variation of h_t and h_b is constraint to a constant value of $h_t + h_b$. Eff_1 and Eff_2, which can be easily calculated from the *HSRs* and are also directly related to Q_1 and Q_2, on the other hand, take also into account the impact of the conductive resistances and, in particular, the impact of R_i, in splitting the heat flux.

Examples of how $\frac{l_b}{l_t}$ and $\frac{h_t}{h_t+h_b}$ affect Θ_{11} are shown in Figure 4.12. Full lines represent cases in which a μbump array is surrounded by underfill, dashed lines are for full homogeneous underfill layer and pointed lines for area μbump array. Different colors represent different sets of system parameters and, on the left plot, the thicknesses of the dies are changed $\left(\frac{h_t}{h_t+h_b} = 0.7\right)$, while on the right plot, the convection coefficient are varied $\left(\frac{l_b}{l_t} = 1\right)$. This picture confirms that, the more room is left for heat spreading below the interface layer, the lower the temperature is in the location of the μbump array.

From this reasoning it is clear that equation (4.6), which just includes Eff_1, is not suitable in case of two sides convection. An adjustable full factorial design of experiments (DOE) has been run to capture the impact of the different parameters

Table 4.3: System parameters used in the DOE to determine the μbumps effect in case of two sides convection.

	h_t	$k_{\mu b,xy}$	$k_{\mu b,z}$	k_{und}	l_t	l_b	l_i	$\tilde{\rho}$
min	1000 W/m^2K	0.3 W/mK	3 W/mK	0.2 W/mK	20 μm	20 μm	6 μm	0.014
max	10000 W/m^2K	10 W/mK	15 W/mK	5 W/mK	400 μm	400 μm	20 μm	1
#values	5	2	2	2	4	4	2	5

on the value of $\Theta_{11,und} - \Theta_{11}$. In particular, attention has been paid to the minimum of this quantity, which is influenced by the two sides convection. A set of parametric, 2D-axisymmetric FEM simulations has been run for different system parameters values and different dimensions of a μbump array placed in the center of the configuration. Note that it is not necessary to consider different μbump layouts for the same $\tilde{\rho}$ value since this effect is already included in the γ parameters. In the DOE, all the system parameters have been included, to confirm the significance of only the ones listed in the previous reasoning. The value of $\Theta_{11,und} - \Theta_{11,c}$, where $\Theta_{11,c}$ is the temperature increase on the top die in the center of the μbump array, has been monitored as output. In order to increase accuracy, more values have been considered between the minimum and the maximum for the parameters that were expected to have a higher impact on the output quantity. More precisely, the list of parameters and the number of values included in the DOE (including the minimum and maximum) is given in Table 4.3. According to equation (4.17), $h_t + h_b$ has been kept fixed at 10000 W/m^2K and the condition $k_{und} < k_{\mu b,xy} < k_{\mu b,z}$ has been imposed.

From the fitting of these results, the μbumps thermal effect has been defined as

$$\mu bEff = Eff_1 + \left(\frac{l_b}{l_t}^{0.12} - 1.57\right)\left(1 - \tilde{\rho}^{-0.1553\left(\frac{l_b}{l_t}\right)^{0.3446}}\right)Eff_2 \tag{4.18}$$

and, from equations (4.6) and (4.18), the temperature profile on the top die can be computed as

$$\Theta_{11} = \Theta_{11,und} - (1-\gamma)\mu bEff. \tag{4.19}$$

The γ coefficients in this equation are the ones obtained by fitting for a configuration with insulation boundary condition on top and whose heat transfer coefficient on the bottom side is $h = h_t + h_b$. This can be done because the interest is not in the real μbumps effect but just in the relative weights.

Equation (4.18) will now be briefly explained. The μbumps effect on the top die for a non-uniform μbump layout is computed as a combination between the μbumps effect detected on the top die, Eff_1, and the one on the bottom die, Eff_2, weighted by the μbump array area ratio, $\tilde{\rho}$, and the ratio between the bottom and the top die thicknesses, $\frac{l_b}{l_t}$. Eff_1 and Eff_2 bring the impact, in splitting the heat flux, of the heat transfer coefficients and of the interface material properties into the equation. If $\tilde{\rho} = 1$, meaning an area array of μbumps, the considered effect should

be just the one detected on the top die and, indeed, in the formula, the bottom effect is multiplied by zero. For the other extreme case, $\tilde{\rho} = 0$, where the interface is completely composed by underfill, the value of equation (4.18) doesn't really matter because the γ value is 1 everywhere and the μbumps effect contribution in computing the temperature via equation (4.19) is nullified. Finally, in case $h_t = 0$, then $Eff_2 = 0$, and equation (4.19) becomes the same as equation (4.6) since $\mu bEff = Eff_1$.

For the intermediate cases, a least square fitting of $\Theta_{11,und} - \Theta_{11,c}$ with respect to the parameters a_i and ε in formula

$$\varepsilon\left[Eff_1 + \left(\frac{l_b}{l_t}^{a_1} + a_2\right)\left(1 - \tilde{\rho}^{a_3\left(\frac{l_b}{l_t}\right)^{a_4}}\right)Eff_2\right] \tag{4.20}$$

has been performed starting from the results obtained in the DOE. The values obtained for the a_i coefficients are the ones reported in equation (4.18), with $R^2_{adj} = 0.983$ and a null p-value. According to equation (4.19), however, $\Theta_{11,und} - \Theta_{11,c} = (\mathbf{1} - \gamma)\mu bEff$. Multiplying $\mu bEff$ by ε allows compensating for the presence of $(1 - \gamma)$ without including the errors coming from the fitting of the γ values. The plots in Figure 4.13 show the comparison between the value of $\Theta_{11,und} - \Theta_{11,c}$ obtained by FEM (line) and the $\mu bEff$ in equation (4.18) multiplied by appropriate ε coefficients (markers) for different values of the system parameters as a function of $\frac{h_t}{h_t+h_b}$. Each plot refers to a particular $\tilde{\rho}$ value and different colors in each plot to different $\frac{l_b}{l_t}$ values. As we can see, the agreement is really good ($\max|err| < 1.2°C$).

The validation of the FTM for two sides convection is finally shown in Figure 4.14 (a) for the same situations illustrated on the left hand side of Figure 4.12. Full lines are for the FEM results while dashed line for the FTM. As we can see, the FTM is able to detect the spreading/constriction effect due to the splitting of the heat flux in two separate sections. The relative error is always less than 1.5%.

Θ_{12} is, then, computed with the same methodology illustrated in the previous Section. This is possible because the difference between Θ_{11} and Θ_{12} is mainly determined by the thermal resistance of the interface layer. Indeed, as Figure 4.14 (b) shows, $\Theta_{11} - \Theta_{12}$, once divided by the heat flux that approximately goes through it $\left(Q_2 \approx \frac{h_b}{h_t+h_b}Q\right)$, is almost independent of the system parameters that are unrelated to the interface layer. This means that, once Θ_{11} is known, Θ_{12} can be computed just knowing the *local* thermal resistance of the interface layer. Equations (4.12) and (4.11) are basically an approximation of this procedure and, therefore, they are used also in case of two sides convections to compute Θ_{12}.

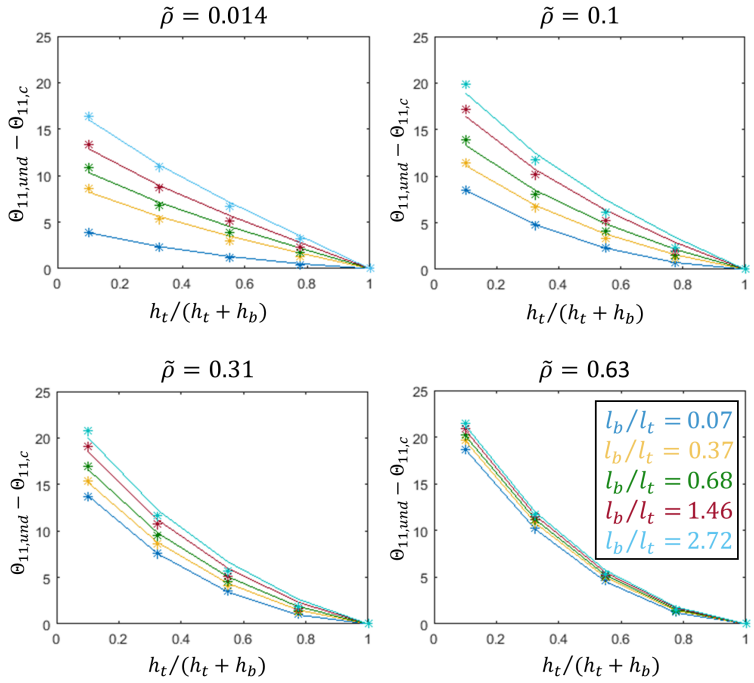


Figure 4.13: $\Theta_{11,und} - \Theta_{11,c}$ obtained by FEM (line) and $\mu bEff$ in equation (4.18), multiplied by appropriate ε coefficients (markers) for different values of the system parameters as a function of $\frac{h_t}{h_t+h_b}$. Each plot refers to a particular $\tilde{\rho}$ value and different colors to different $\frac{l_b}{l_t}$ values.

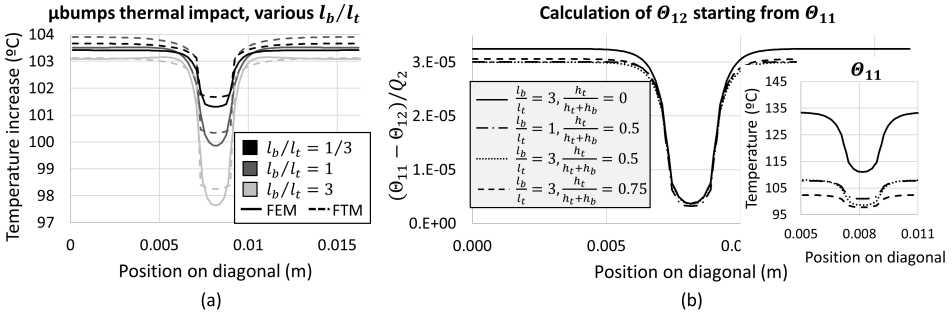


Figure 4.14: (a): Validation of the calculation of Θ_{11} for two sides convection for the same cases on the left hand side of Figure 4.12 (full lines= FEM, dashed lines=FTM). (b) Validation of the methodology to determine Θ_{12} starting from Θ_{11}.

4.3.3 Uniform power on both dies, convection from both sides

In case power is dissipated on the bottom die, Θ_{22} and Θ_{21} are computed using the same methodology illustrated in the previous Sections but assuming the stack to be flipped upside-down. When power is dissipated on both dies, the effect due to the heating of the top die and of the bottom die are computed separately and, then, exploiting the superposition principle, results referring to the same layer are summed up

$$\Theta_1 = \Theta_{11} + \Theta_{21}, \qquad \Theta_2 = \Theta_{12} + \Theta_{22}. \tag{4.21}$$

4.3.4 Non-uniform power on both dies, convection from both sides

Up to now, uniform power maps have been considered: in case heat is non-uniformly dissipated, the specific power maps are substituted in all the previous steps, except in the fittings, to the uniform ones. If, on the one hand, the computation of the parameters for uniform power dissipation allows the easy extension of the model to various power dissipation scenarios, it creates, on the other hand, artificial effects especially at those positions where no power is dissipated and a transition between die-die interface materials is present. To overcome this problem, the μbumps effect in equation (4.19) in locations where little power is dissipated is reduced, by means of a filtering procedure similar to the one illustrated in Figure 4.5, taking into account how much power is dissipated in that position and around it.

4.3.5 Flowchart of the FTM algorithm

Figure 4.15 shows the flowchart reporting the steps needed to implement the algorithm presented in this Chapter. The block listing the initial inputs (first block in Figure 3.10 for instance) is not reported here but it is the same as the one in the previous flowcharts. This chart schematically shows how to, once the four HSRs are computed by FEM (gray block), combine them according to the specific μbump layout, material properties, geometry and BCs in order to include both the local and the global thermal impact of the interface heterogeneity.

4.4 Results and comparison with FEM simulations

In this Section, the results obtained by this extended FTM are validated with respect to FEM models of exactly the same structure (package included as boundary conditions and not fully modeled). In the first column of Figure 4.17 the μbumps

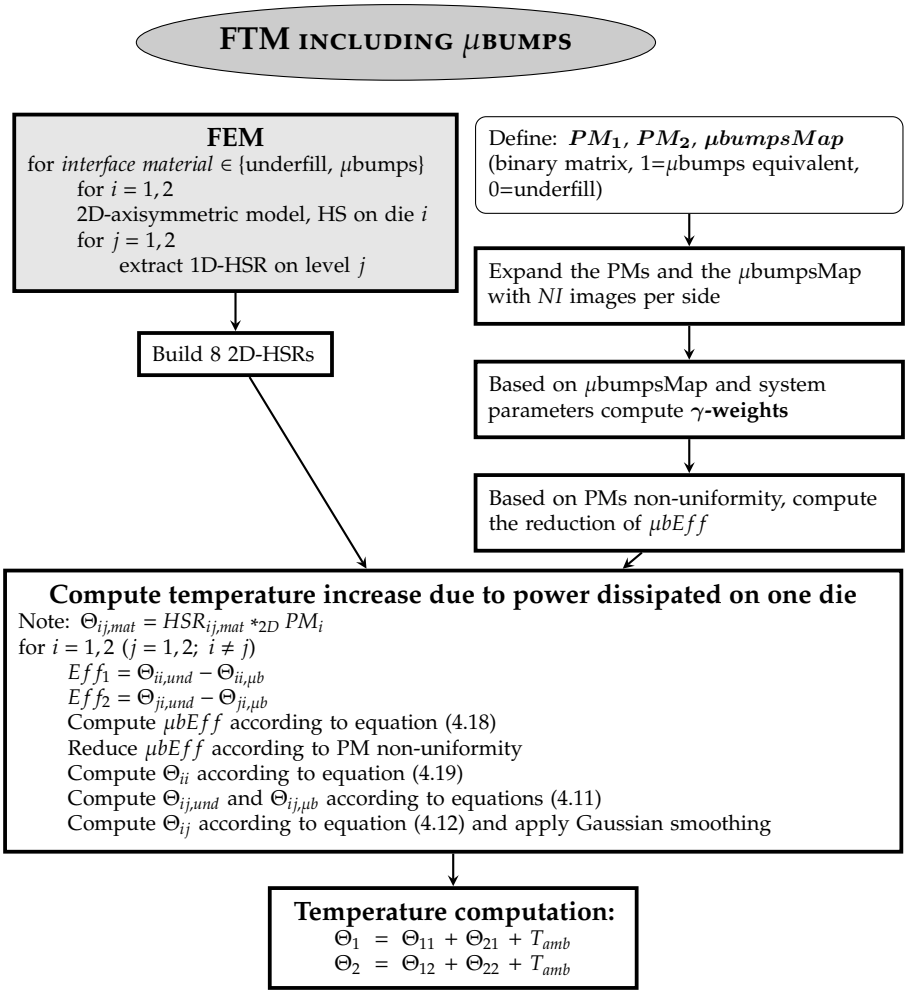


Figure 4.15: Flowchart representing the algorithm implemented for the steady state fast thermal modeling of two dies stacks in F2F configuration including the local and global thermal impact of specific μbump arrays.

map and the PMs are shown while the full temperature profiles obtained by FTM on the two dies are reported on the last row together with the percentage errors (computed as $(T_{FEM} - T_{FTM})/(T_{FEM} - T_{Amb})$). The other plots in the Figure refer to the cross sections along the diagonal and the anti-diagonal of the top (first row) and bottom (second row) die. Blue lines represent the FTM results, red lines the FEM solution while black and green curves denote, respectively, the solution for analogous structures in which homogeneous underfill or homogeneous μbumps equivalent materials are assumed in the interface. The parameters used for this

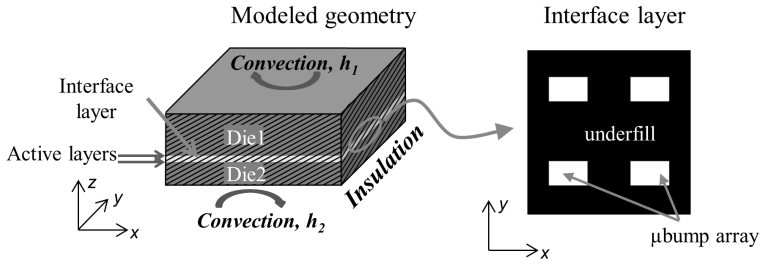


Figure 4.16: FEM setup used to validate the FTM including specific μbump arrays (cf. Figure 4.17). The values of the parameters are reported in Table 4.4 while the dissipated PMs in Figure 4.17.

Table 4.4: System parameters used to obtain the results in Figure 4.17.

h_t	h_b	$k_{\mu b,xy}$	$k_{\mu b,z}$	k_{und}	k_{Si}	
950 W/m^2K	50 W/m^2K	0.6 W/mK	4.2 W/mK	0.4 W/mK	120 W/mK	
l_t	l_b	l_i	cs	$\bar{h}$	T_{Amb}	# images
150 μm	150 μm	13 μm	8.16 mm	120 μm	25 °C	7

simulation are listed in Table 4.4 while a sketch of the modeled situation is reported in Figure 4.16.

Good agreement is achieved between the FTM and the FEM model: solutions considering uniform horizontal material layers give higher errors especially in areas where, in fact, the other material is used. This means that the developed FTM is able to detect both the local and the global thermal impact of the particular μbump layout in the die-die interface layer. This is important in case specific temperature limitations must be fulfilled while designing the 3D-IC. Knowing how large the thermal improvement will be by introducing more expensive structures, such as thermal dummy μbumps, can help in developing a better and more efficient design.

In this particular case, the relative difference between FTM and FEM prediction is less than 1.3% while in all the considered test cases it has been evaluated to be less than 2%. The overall percentage error profiles, presented in the last row of Figure 4.17, show that the maximum error is achieved in the critical zones, where discontinuity (either in power dissipation or in interface material) occurs, while, elsewhere, it is close to 0%.

As well as accuracy, the computational speed is another important property of a FTM. In this case, on a standard PC, the temperature profiles are obtained in around 5 sec, which is already 20 times faster than using FEM (just the running time is considered). These results have been obtained using just one thread. However,

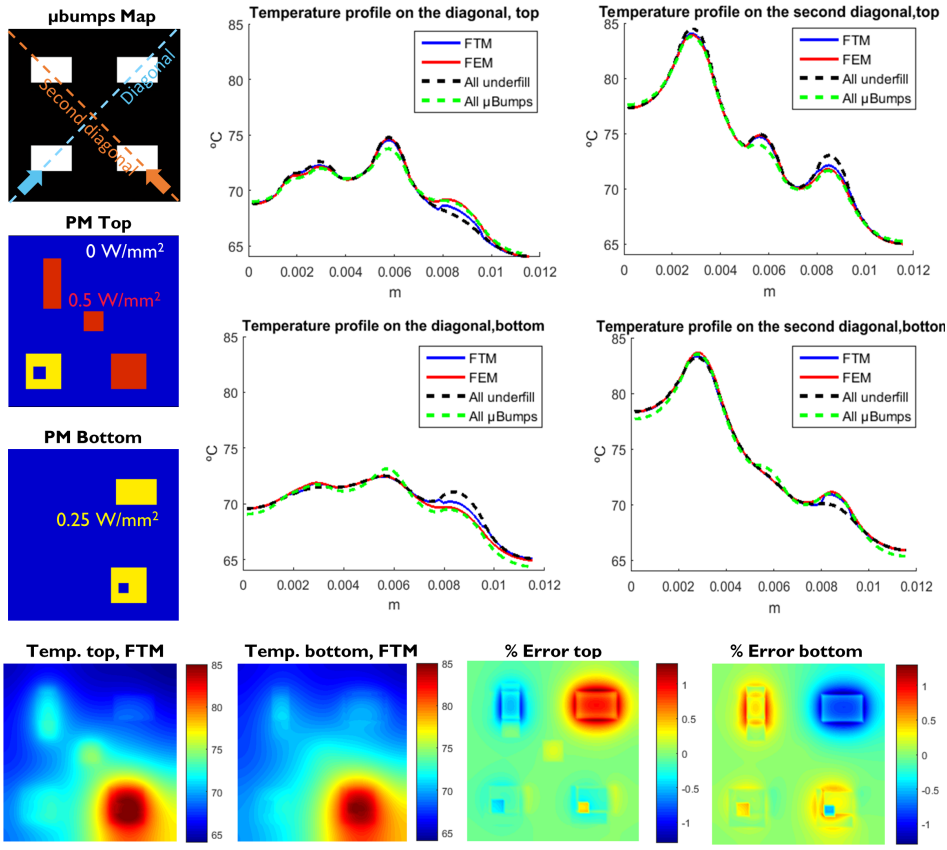


Figure 4.17: Temperature profiles on top and bottom die obtained using the illustrated power maps and μbumps map. The anti-diagonal and diagonal cross sections of the FTM result (blue), of the FEM model (red), of a structure assuming underfill everywhere (black) and μbumps everywhere (green) are shown. The last row presents the whole temperature profiles on the top and bottom die and the relative percentage errors.

the FTM model can be easily run in parallel, since the temperature profiles due to power dissipation on the top die (Θ_{11} and Θ_{12}) and on the bottom die (Θ_{21} and Θ_{22}) are computed separately. Moreover, different configurations can be easily tested by changing the input parameters reported in Table 4.4 or the entries in the power maps and in the μbumps map. In particular, if just the power maps and/or the μbumps map change, since the HSRs are already known, no further 2D-axisymmetric FEM simulations are needed.

Figure 4.18 show the importance of considering specific μbump layouts in the thermal analysis. Different μbump layouts (left side of the Figure) result, indeed,

Table 4.5: System parameters used to obtain the results in Figure 4.18.

h_t	h_b	$k_{\mu b,xy}$	$k_{\mu b,z}$	k_{und}	k_{Si}	Power (top)
insulation	8000 W/m^2K	0.75 W/mK	6 W/mK	0.5 W/mK	120 W/mK	66.6W
l_t	l_b	l_i	cs	$\bar{h}$	T_{Amb}	# images
200 μm	50 μm	13 μm	8.16 mm	120 μm	0 °C	5

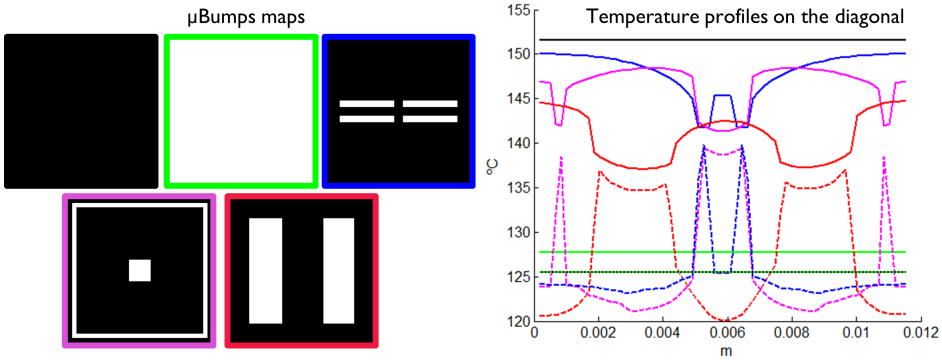


Figure 4.18: Diagonal cross-sections (right) of the temperature profiles obtained dissipating uniform power on the top die of stacks whose μbump layouts are shown on the left (black=underfill, white=μbump array). Full lines refer to top die temperatures while dashed lines to bottom die temperatures; insulation is assumed on the top boundary.

in different temperature profiles (right side of the Figure), which this FTM is able to capture. The cross sections, represented in the rightmost graph, refer to the temperature on the diagonal of the top die (full lines) and of the bottom die (dashed lines) of a structure in which uniform power is dissipated on the top die and heat is convectively removed just from the bottom side. The modeled geometry is the same as the one shown in Figure 4.16 while the considered parameters are listed in Table 4.5. Despite uniform heat dissipation, non-constant temperature profiles are computed by the FTM, proving that the model is able to deal with material non-homogeneity. Moreover, from these plots it is clear that the effect of the μbump array is both local and global: the temperature profiles achieved considering specific non-uniform μbump layouts differ, indeed, in *each* location from the ones obtained considering uniform interface materials. They are, for example, different from $T_{11,und}$ everywhere, not just close to the locations where μbumps are placed. This is why a global approach, which starts with a fitting procedure that takes into account the whole system, is necessary. It's worth to stress that, to check the global thermal impact of different μbump layouts for fixed system parameters, as in this analyzed case, just the entries in the binary μbumps map need to be changed before re-running the model (less than 5 sec

computational time).

4.5 Summary

In this Chapter, a correction methodology to account for the steady state thermal impact of specific material heterogeneity (μbump layout) in the interface layer of two dies stacks in face-to-face configuration has been presented. Since in presence of material heterogeneity the heat path depends on the horizontal location where power is dissipated, the convolution based FTM, which requires the heat path to be independent from this position, is not applicable anymore. As a consequence, a correction methodology has been developed and described. Starting from a simplified case in which uniform power is dissipated on one die and the heat is convectively removed from the bottom side, it has been shown that fitting of various FEM results can lead to an accurate correction model. Subsequently, the method has been further generalized allowing structures with cooling on both sides and non-uniform heat generation in both chips.

The final model, therefore, allows assessing the thermal impact of different system parameters (heat transfer coefficients, dies and interface thicknesses, material properties) as well as of the placement of the power dissipation areas and of the μbump arrays. User defined μbump layouts can, indeed, be considered in the die-die interface layer and their global thermal effect is accurately captured with a maximum error less than 2% with respect to FEM models of exactly the same stack configurations.

Another important aspect that a FTM should fulfill is computational efficiency: this code results to be around 20 times faster than analogous FEM models (only the running time is accounted for and parallelization of the FTM is not exploited). Moreover, thanks to its ease of use, the effect of various parameters can be checked just changing input numbers or entries in matrices. Thus, a useful early design phase tool is established. The distinctive features of the FTM presented in this Chapter are, therefore, its ability to provide the complete horizontal temperature profiles together with the combination of ease of use, speed, ability in detecting the thermal impact of heterogeneous die-die interface material and accuracy.

A possible application of this methodology will be discussed in Section 9.4.2.The example will show, in particular, that the maximum temperature can be kept below a user defined critical limit by appropriately placing μbump arrays. The amount of μbumps and their positioning can be forecast by the algorithm in such a way that no unnecessary thermal dummy μbumps, which would anyhow increase the cost of the device, are included in the design.

Chapter 5

Package Thermal Spreading

5.1 Introduction

The limitation of the convolution based FTM concerning the modeling of stacked structures in which each horizontal layer is made of just *one* homogeneous material is not the only one related to the position independence requirement of the HSRs. Another limitation is that all the layers in a stacked structure need to have the *same* horizontal area. In more realistic scenarios, however, the die stacks are enclosed in packages, which are placed on printed circuit boards, and are cooled down by application dependent cooling solutions (heat sink, fan, liquid cooling, . . .). All these parts are normally much larger than the die stack itself, allowing for lateral heat spreading and temperature reduction (cf. Section 1.3.1). The convolution based FTM for 3D-stacks of finite dimensions (it will be referred to as *stack FTM* from now on) is not able to directly include this effect, since the spreading resistance is position dependent. The spreading effect is, indeed, larger close to the corners of the stack, where more surrounding passive material is present, than in the center of the dies. Moreover, in some package configurations, the die stack is surrounded by other passive materials (epoxy mold compound, for instance), meaning that more materials are present on the same horizontal layer.

Stack structures, with finite horizontal dimensions, have been considered up to now while validating the FTM with respect to FEM. This has been done in order to focus the validation on the modeling of the internal conduction rather than on the impact of the BCs. The effect of the rest of the package was included through equivalent convective BCs, with constant heat transfer coefficients, applied to the top and bottom sides of the die stack. The low error computed for these stack configurations demonstrates that the FTM is able to properly predict the temperature distribution in this setup. However, as already mentioned, the actual geometry of the package may have a significant impact on the temperature profiles

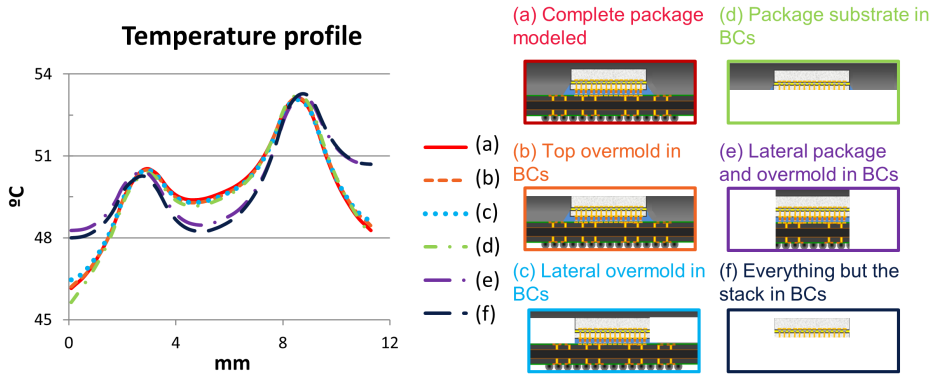


Figure 5.1: Package impact analysis using FEM. The graph shows the diagonal of the temperature profile on the bottom die obtained by including different parts of the package in the model rather than including them in the top and bottom uniform convective boundary conditions.

in real applications (cf. Figure 5.1 (a) for package configuration and Figure 5.1 (f) for stack configuration). This means that, to provide a proper estimation of the temperature profiles, the *whole* structure should be taken into account. However, due to the difference in the length and time scales of the different parts constituting the full packaged device, these parts cannot be considered all in the same way and with the same resolution. Depending on the level (die, package, system, . . .) of interest of the simulation itself, some of these parts are not explicitly and accurately modeled but they are included in the outside environment. Their effect is, then, mimicked by applying case dependent boundary conditions, which are normally convective or insulating.

The application of these BCs can provide a satisfactory estimation of the environmental thermal impact for certain configurations in the steady state regime but, for transient simulations, the capacitive effect of the un-modelled parts cannot be included. This effect is related to the accumulation and release of heat during chip activity and it has a significant impact on the time constant of the problem. It may be, therefore, important to include the information about the package in the FTM simulations. It would allow, indeed, to account for the extra thermal spreading and, in case of transient simulations, also for the extra capacitance that is not included in the stack configuration.

In this Chapter, a correction methodology, to be applied on top of the temperature results obtained by the stack FTM, is presented both for the steady state and the transient regime. It can be considered as a multi-scale strategy whose core is constituted by the highly resolved, convolution based algorithm that allows to compute the temperature increase due to a generic, time varying, power map in all the dies of the stack configuration. On top of this, the package spreading and

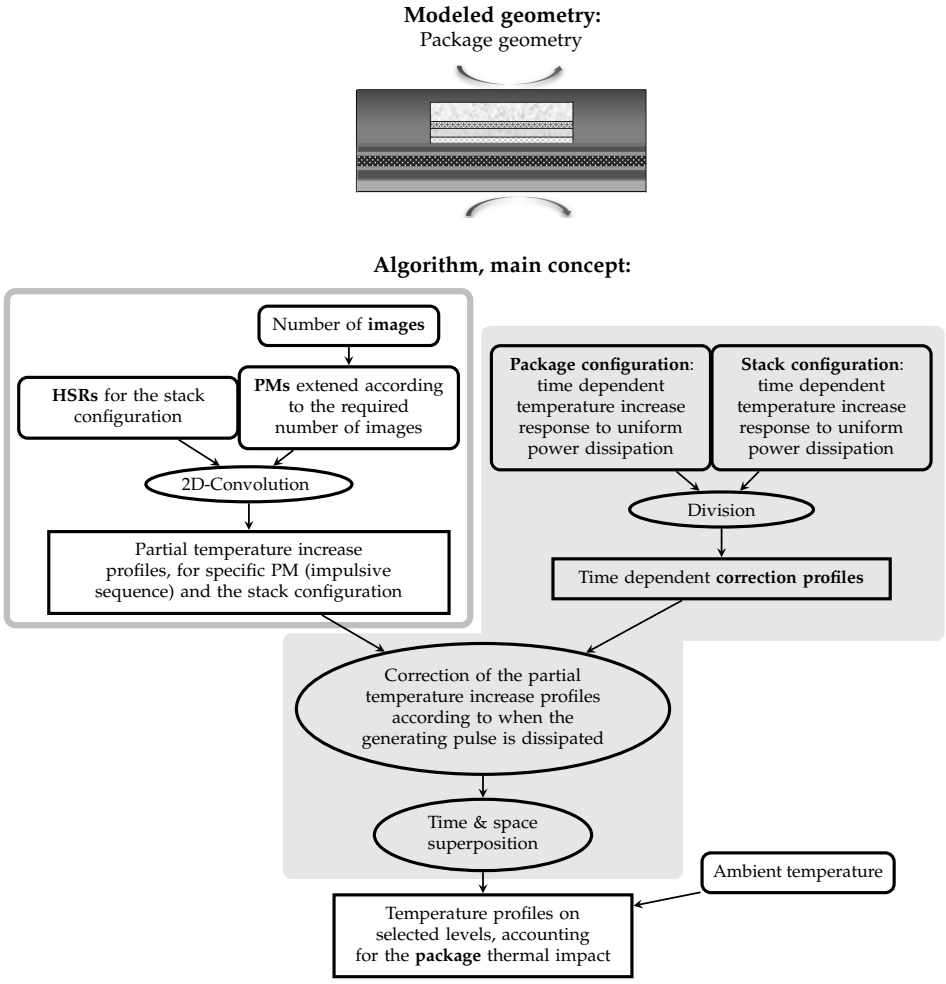


Figure 5.2: Modeled geometry and main concept of the algorithm described in this Chapter. The section specifically introduced and discussed in this Chapter is highlighted.

capacitive effect is included via correction profiles. They are based on the ratio between the steady state or time dependent thermal responses of the package and of the stack configurations to uniform, impulsive, power dissipation. The main concept of the algorithm, which is described in details in the following of this Chapter, is shown in the flowchart in Figure 5.2. In the chart, the parts that are added with respect to the FTM presented in Chapter 3, are highlighted in gray. It is worth to stress now that, for the transient regime, a modification of the algorithm developed to compute the temperature profiles in the corresponding stack configuration has

to be implemented. While, for a simple stack configuration without the package thermal impact, a 3D-convolution algorithm can be implemented to obtain the temperature profiles, if the time dependent package thermal impact has to be taken into account, a 2D-convolution approach with subsequent time superposition is needed. This causes an increase in computational time. Nevertheless, the package FTM remains significantly faster than analogous models based on FEM.

The model has been validated with respect to FEM results for the full package, showing good accuracy (cf. Sections 5.3.6 and 5.4.5). Moreover, an error metric, to estimate a priori the need of the package correction on top of the convolution based approach, has been developed (cf. Paragraph *"Error metric "* in Section 5.4.3). Finally, alternative but similar algorithms, which place themselves in between the corrected and the uncorrected approach, both from an accuracy and from a computational time point of view, are shortly presented in this Chapter (cf. Paragraph *"Alternative package correction approaches"* in Section 5.4.5).

5.2 Impact of the package on the thermal modeling results

A full 3D FEM steady state study has been performed, and published in [60], to assess the impact of the division of the package into a region of interest, in which the modeling is performed, and a second region, which is replaced by equivalent, constant BCs acting on the region of interest. The analysis is performed for a specific case in which a package with high thermal resistance is considered. More precisely, a $8 \times 8mm^2$ die stack, overmolded using epoxy mold compound and packaged on a $14 \times 14mm^2$ substrate, is considered. For more information about the FEM model itself, please refer to Appendix A and Section 7.4. For different package configurations the results may be different. Figure 5.1 shows the diagonal cross sections of the temperature increases on the bottom die obtained, under steady state conditions, including different parts of the package in the modeled region. The Figure shows that there is a significant impact of the full package on the temperature profiles but also that not all the sections of this more complex structure play a role in deviating the temperature profile from the one obtained considering just the stack configuration.

The full red line (case (a)) represents the case in which the complete package is modeled. Materials around the die stack are consecutively removed and the BCs are adapted so that the same $\max(T)$ is maintained. In this way, the impact of each specific part on the final temperature profile can be identify. The blue dashed line (case (f)), for example, is used for the situation in which the structure considered in the stack FTM described in Chapter 3, i.e. the stack configuration, is modeled. The thermal effect of the package is clearly visible in the comparison. Considering a structure as the one the stack FTM can deal with, results in a maximum error

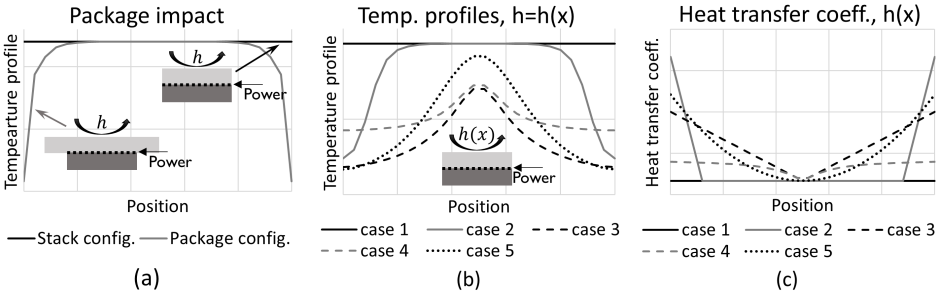


Figure 5.3: Mimicking the package thermal impact by position dependent heat transfer coefficient (uniform power dissipation). (a): Temperature profiles for the stack (black) and the package geometry (gray) assuming constant h. (b): Temperature profiles obtained considering position dependent heat transfer coefficients (h). The dependency of h on x is shown in (c).

around 5.5% in the chip corners. This is not related to the modeling *strategy* itself (all the results are obtained by FEM) but to the modeled *structure*. In case just the part of the package around the die stack is removed maintaining the same cooling area on top and bottom, case (c), the difference with respect to the modeling of the complete package structure is negligible. This shows that the application of insulating boundary conditions on the lateral sides of the stack is not responsible for this difference. Instead, the main reason for the thermal difference in including or not the package in the modeled region is identified in the total surface available for cooling, i.e. in the spreading effect on top and bottom of the stack. As soon as this area is reduced (cases (e), (f)), the difference becomes more significant. Even if the *constant* heat transfer coefficients are scaled, as in this example, to account for the reduction of the area available for cooling, the final shapes of the temperature profiles are different than in case the complete package is modeled.

A possible solution is the use of non-uniform heat transfer coefficients for the BCs applied on top and bottom of the die stack. The variation of the spreading resistance experienced in the different horizontal positions of the stack can, indeed, be mimicked by position dependent BCs. This would allow to appropriately model the heat transfer phenomenon experienced by the packaged 3D-IC by just modeling the die stack. Figure 5.3 (a) shows the steady state temperature profiles obtained for a package configuration with a Cu heat spreader on top of the silicon die (gray) and for the corresponding stack configuration (black). In both models the Cu layer is present; the difference is that, in the first case, this layer is larger than the die itself while, in the second case, it is as large as the die. This last geometry is the kind of structure that has been considered in the FTM up to now. Sketches of the considered package and stack configurations are reported in Figure 5.3 (a). Uniform power is dissipated on top of the die and the same, constant heat transfer coefficient is applied on the top boundary of the two configurations. All

the other boundaries are assumed to be insulated. As the Figure shows, the effect of the package is to reduce the temperature close to the edges of the stack. This is because the package has a larger area than the die stack and, as a consequence, it allows more heat spreading close to the edges. This kind of temperature profile can be mimicked by assuming a position dependent heat transfer coefficient (h) and a stack configuration (all layers have the same horizontal area). Figure 5.3 (b) shows, indeed, different temperature profiles that can be obtained for the same stack configuration but letting the h vary with the position along the die (cf. Figure 5.3 (c)). A proper definition of $h(x, y)$ could, therefore, include the thermal impact of the package maintaining the *geometry* manageable by the stack FTM. Unfortunately, also this approach cannot be implemented in a convolution based modeling strategy: the temperature response of the system would, indeed, be dependent on the position where power is dissipated, which violates one of the basic assumptions of the convolution based FTM. It is worthed to note that all the results in Figure 5.3 refer to 2D models; the conclusions of this analysis remain, however, valid also for 3D structures.

5.3 Steady state regime

5.3.1 Previous work

Hériz et al. proposed in [37], for the steady state regime, an error reduction technique to include the thermal impact of the package in their convolution based FTM methodology, which had been developed for the stack configuration (cf. Figure 5.4 for the terminology). Their approach is based on the computation of the *intrinsic error*, due to the difference in geometry, between the temperature profiles obtained for the stack and the package configurations. It is computed by comparing the system responses to uniform power dissipation in the two different cases. More precisely, the temperature increase profile for the stack geometry, $\Theta_{stack,unif}$, is computed by the stack FTM while the one related to the package geometry, $\Theta_{pack,unif}$, by means of FEM. To obtain the temperature profile, $\Theta_{FTM,pack}$, which includes the package thermal impact for a non-uniform PM, the temperature profile, $\Theta_{FTM,stack}$, obtained by the stack FTM for the same PM, is corrected as

$$\Theta_{FTM,pack} = \frac{\Theta_{FTM,stack}}{1 + E_r} \tag{5.1}$$

where

$$E_r = \frac{\Theta_{stack,unif} - \Theta_{pack,unif}}{\Theta_{pack,unif}}. \tag{5.2}$$

Rearranging of equation (5.1) leads to

$$\Theta_{FTM,pack} = \Theta_{FTM,stack} \cdot \bar{C} \tag{5.3}$$

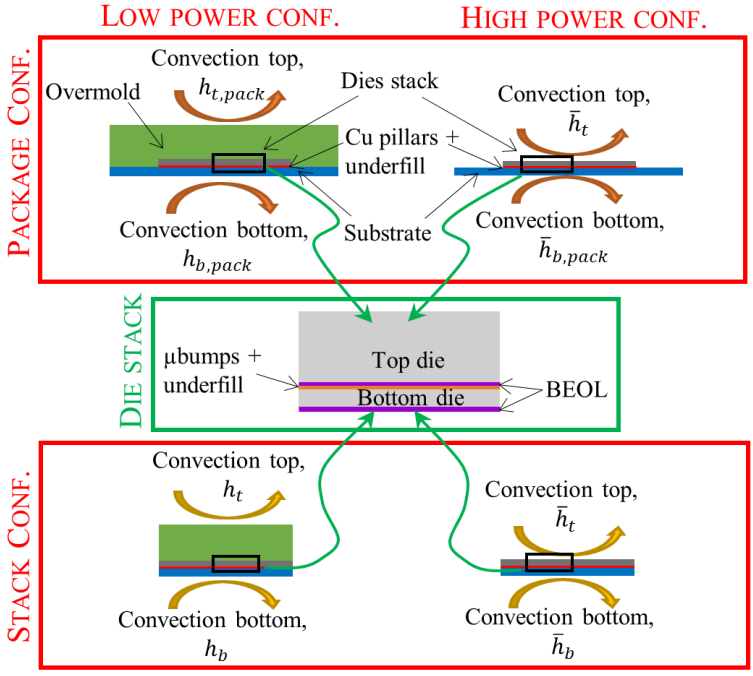


Figure 5.4: Illustration of the nomenclature and of the package configurations, for the low power (LP), on the left, and the high power (HP), on the right, cases presented in this Chapter. What package configuration, die stack and stack configuration refer to, is, respectively, illustrated on the top, center and bottom part of the Figure.

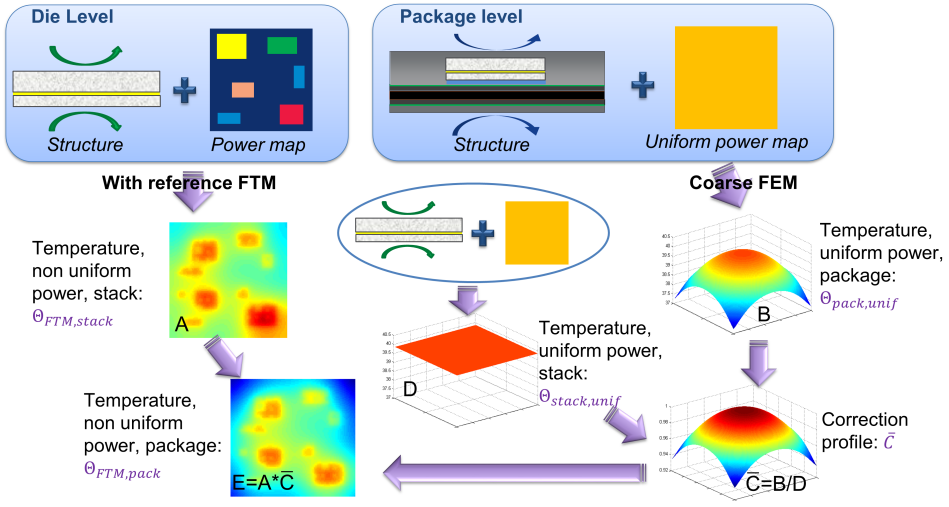


Figure 5.5: Illustration of the methodology to include the package thermal impact in steady state simulations.

where

$$\bar{C} = \frac{\Theta_{pack,unif}}{\Theta_{stack,unif}} \tag{5.4}$$

is a correction factor carrying the information about the spreading effect that is intrinsically neglected in the stack FTM. An illustration of this methodology for steady state is shown in Figure 5.5.

Two main package configurations are considered in this thesis: a *low power* (LP) configuration (left hand side of Figure 5.4) and a *high power* (HP) configuration (right hand side of Figure 5.4). In both cases, the die stack is attached to a substrate by a layer of Cu pillars and underfill, for which equivalent material properties are used. In the low power configuration the die stack is overmolded while, in case of high power, the cooling is directly applied on the backside of the top chip. Equivalent properties are also used for the μbumps-underfill layer in between the two dies.

5.3.2 Physical base

In this Section, the physics, on which this steady state correction methodology is based, is explained for a simple 2D case. Newton law of cooling states that

$$Q = hA\Theta_s \tag{5.5}$$

where Q is the total dissipated power, Θ_s is the temperature difference between the surface of the object and the ambient and h is the heat transfer coefficient, which can be a constant value or position dependent. Let's assume a stack configuration with uniform power dissipation and a one-side cooling with a constant value of h. Making use of the RC-network approach, the conduction and the convection thermal resistances from the power dissipation level to the ambient can be combined and an equivalent heat transfer coefficient $\tilde{h}_{stack}$, mimicking the whole heat path, can be defined. This means that

$$\Theta_{stack,unif} = Q\frac{1}{\tilde{h}_{stack}A}. \tag{5.6}$$

According to the explanation in Section 5.2, the thermal behavior of a package configuration can be mimicked by assuming a stack configuration and a position dependent heat transfer coefficient, $h_{pack}(x, y)$. As a consequence, for the steady state regime,

$$\Theta_{pack,unif} = Q\frac{1}{\tilde{h}_{pack}(x, y)A}. \tag{5.7}$$

Let's now assume a generic, non uniform power dissipation in a package configuration: the die is discretized and a value for the dissipated power and for the equivalent heat transfer coefficient is assigned to each cell. In particular,

the values of $\tilde{h}_{pack}(x,y)$ are the ones calculated for uniform and constant power dissipation scenario. The temperature can, therefore, be approximated as

$$\begin{aligned}\Theta_{FTM,pack}(x,y) &\approx Q(x,y)\frac{1}{\tilde{h}_{pack}(x,y)A} = Q(x,y)\frac{1}{\tilde{h}_{stack}A}\frac{\tilde{h}_{stack}A}{\tilde{h}_{pack}(x,y)A} \\ &\approx \Theta_{FTM,stack}(x,y)\frac{\Theta_{pack,unif}(x,y)}{\Theta_{stack,unif}} = \Theta_{FTM,stack}(x,y)\cdot\bar{C}.\end{aligned} \tag{5.8}$$

The approximation symbols come from the fact that the lateral heat spreading *inside* the die stack, due to non uniform power dissipation, is initially not accounted for (package configuration) and it is then reintroduced in a second stage (stack configuration). The correction procedure to include the package thermal impact on top of the results obtained by the stack FTM has been built assuming that the difference in this lateral spreading between the two configurations is small and that the dependence of the $\tilde{h}(x,y)$ coefficients on the non-uniformity of the power map can be neglected.

5.3.3 Bottleneck of Hériz's methodology and possible solutions

The main bottleneck, from a computational time point of view, of the methodology proposed by Hériz et al. in [37] is that, for each specific configuration (i.e. geometry, BCs and materials), a full FEM simulation for uniform power dissipation is needed. For this reason, various approaches have been investigated to increase the computational speed by avoiding the use of FEM.

Analytical solution

As reported in Section 1.3.2, under particular conditions, analytical solutions are available in literature to compute the temperature profiles. They are based on series expansions, meaning that they may require high computational time, and they are only valid for simplified situations, meaning that some of the essential features cannot be modeled [30]. One of these constraints is the assumption that power is dissipated on one boundary of the stack and that the heat is removed from the other side. This means that on one side of the power source no material is considered and that convection from two sides is not allowed. This latter constraint can be acceptable in situations where the heat path can be considered one directional (e.g. a highly efficient cooling solution on top of the stack) but not in general.

Moreover, in a real situation, the power is dissipated inside the stack and solid material is present all around it. Assuming, for example, that a heat spreader/heat sink is applied on top of the stack, as shown on the right hand side of Figure 5.6

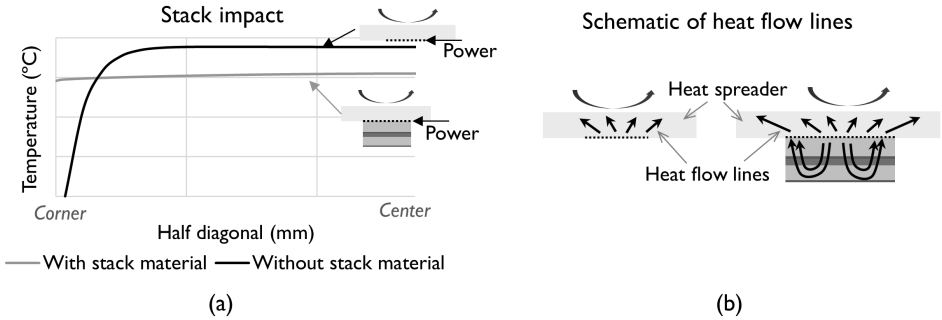


Figure 5.6: (a): Temperature profiles for uniform power dissipation with (gray) and without (black) the influence of the die stack. Results refer to the half diagonal of the die in a 3D model. The black curve is the one that can be obtained analytically. (b): Schematic of the heat flow lines in the two cases.

(b), and that the heat is mainly removed from that side of the device, the presence of the die stack below the power dissipation level has a strong impact on the final temperature profile. This happens even if the stack is *outside* the primary heat path. This effect is shown in Figure 5.6 (a) where the gray curve indicates the temperature profile, from the chip corner to the center, obtained by FEM for uniform power dissipation considering the stack below the heat source, while the black curve indicates the temperature profile obtained by the analytical solution [30]. In this latter case, the heat spreading due to the larger area of the heat spreader/heat sink with respect to the power dissipation surface, which corresponds to the stack surface, is accounted for but the effect of the underlying stack is neglected. In both cases, the results refer to 3D models.

As it is clearly visible from Figure 5.6, the presence of the stack modifies the heat path since, through the stack, part of the heat dissipated in the center moves towards the edges, which are colder, because of the spreading effect of the larger layer above the heat source, before being convectively transferred to the ambient (Figure 5.6 (b)). The stack acts, therefore, as a sort of box where heat is redistributed and, as a consequence, the temperature profile becomes flatter.

If we want to stick to a modeled configuration without the die stack, which can be managed by an analytical approach, the impact of the stack can be seen as a *power redistribution effect*. This means that the dissipated power should not be considered uniform inside the active region but a higher value should be considered close to the edges and a lower one in the center. However, the way in which power is redistributed depends on all the geometric and material parameters and a way to translate the stack effect into power redistribution has not been found.

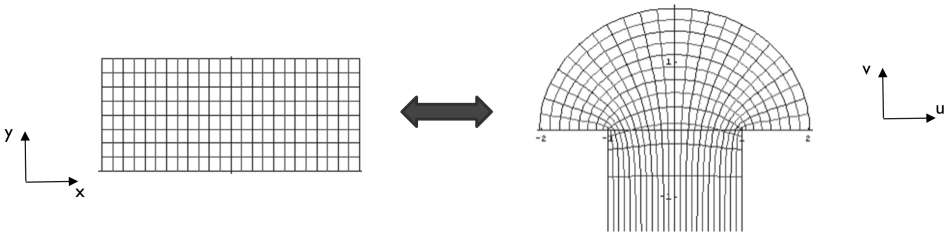


Figure 5.7: Proposed conformal map transforming the temperature profile obtained for a stack configuration into a geometry with a larger top layer (from [64]).

Conformal mapping

Conformal mapping is a technique used to map a complex geometry, where a PDE is difficult to be solved, to a simpler one, where the solution is known (and vice versa) [42]. The main property of this transformation is that it locally preserves angles, meaning that, for the steady state heat conduction equation, the heat flux lines and the temperature contours remain perpendicular to each other while changing the geometry. The idea is to use a geometry transformation as the one illustrated in Figure 5.7 to map the heat flow and the temperature contour lines from a simple geometry (left hand side), to a more complex one (right hand side). The simple geometry resembles the stack configuration, which can be handled by the FTM and in which all the layers are assumed to have the same horizontal area, while the more complex geometry can be a representation of a situation in which a heat sink or overmold material is placed on top of the die stack. As demonstrated in Figure 5.1, indeed, the overmold material on the lateral side of the stack doesn't have a significant impact on the final temperature profile and it can, therefore, be neglected. The equation characterizing this transformation is [64]

$$w = u + iv = \frac{2}{\pi}\sqrt{(y\cos x + iy\sin x)^2 - 1} + \frac{2}{\pi}\arcsin\left(\frac{1}{y\cos x + iy\sin x}\right) \qquad (5.9)$$

with $y > 0$ and $0 < x \leq \pi$.

On top of being a 2D transformation while the stack geometry is three dimensional, there are other fundamental differences between the 2D cross section of the geometry of a stack with a larger overmold area (or heat sink) on top of it and the one reported on the right hand side of Figure 5.7. Some information is, indeed, missing in this last geometry. In particular:

- the die stack is considered infinitely thick and its horizontal dimension is fixed to 2;
- the overmold (or heat sink) on top of the stack is considered infinitely large and thick;

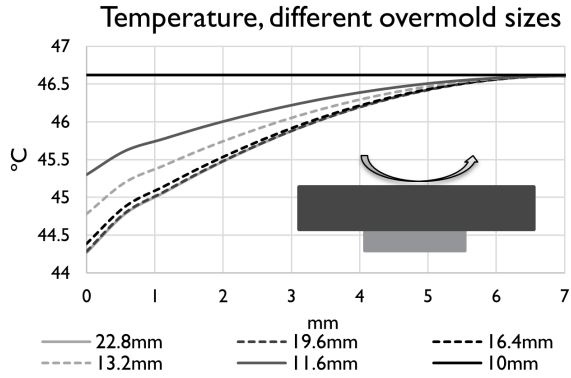


Figure 5.8: Diagonal of the temperature profiles at die level for different overmold sizes (in mm). The curves are translated to have the same maximum and the die stack is $10 \times 10\ mm^2$.

- the difference in the constituent materials is not included as well as the impact of the applied BCs. This means that the transformation is not influenced by these parameters;
- a level where the temperature profile has to be extracted needs to be selected and related to the real geometry.

The analysis of the impact of the real dimension of the overmold on the temperature profiles has been performed through some FEM simulations. A $10 \times 10\ mm^2$ die stack is considered and the size of the overmold is varied to check the validity of the assumption of having an infinitely large top layer. The half diagonals of the temperature profiles obtained on the power dissipation level are shown in Figure 5.8. The heat transfer coefficients have been adapted so that the temperature in the center is the same for all the cases. The graph shows that, for this particular setup, the size of the overmold has an impact on the obtained temperature profile if it is smaller than $16.4 \times 16.4\ mm^2$: for all larger configurations, the temperature profiles remain the same. Typically, the floor-plan area of the substrate (and of the overmold) is at least two times larger than the one of the die stack itself allowing, in principle, the use of the conformal mapping approach with an infinite large overmold. The dimension at which the size of the overmold becomes irrelevant has, however, also a dependence on the boundary conditions and on the material properties.

The main issue regarding the application of this approach concerns the thickness of the stack. Opposite to the analytical approach, in which the die stack cannot be included, in the conformal mapping transformation the stack is assumed to be infinitely thick. Also this assumption has a significant impact on the obtained temperature profiles. A FEM study has been performed to analyze the impact of

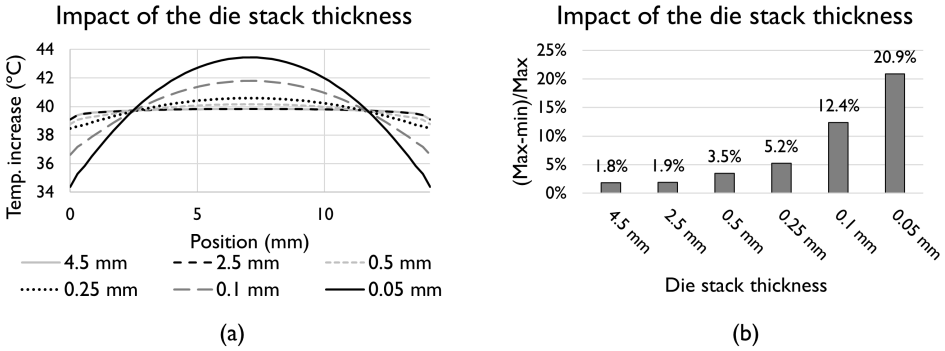


Figure 5.9: Impact of the thickness of the die stack on the temperature profiles in a package configuration. (a) Temperature profiles on the diagonal of the stack. (b): Percentage difference between the maximum and minimum values of the obtained temperature profiles.

the thickness of the die stack on the temperature. In the model, a larger overmold is placed on top of the stack, insulation is assumed on the bottom of the stack and uniform power is dissipated between the stack and the overmold material. In Figure 5.9 (a) the diagonal of the temperature profiles on the power dissipation level are reported for different thicknesses of the die stack. Figure 5.9 (b) reports the percentage difference between the maximum and the minimum of the obtained temperature profiles. From the two graphs it is clear that the assumption of an infinitely long stack highly affects the temperature response of the system. In case of stacks of few dies and/or if the dies are thinned down, indeed, the difference between the temperature in the center and in the corner can be as large as 20%, while the conformal mapping strategy would return a value less than 2%. This means that also the conformal mapping approach doesn't seem to be able to properly estimate $\Theta_{pack,unif}$.

A significant property that emerged during this analysis is that, if the temperature profiles obtained for different configurations are scaled so that both the maximum and the minimum values coincide, all the curves (except the flat one for which the size of the overmold is equal to the one of the stack) coincide (cf. Figure 5.10). It seems, therefore, that any profile can be retrieved by just scaling an *original surface*, once the minimum and the maximum temperature for the specific case are known. This is the strategy that has been developed in this work and that is explained in the next Subsection.

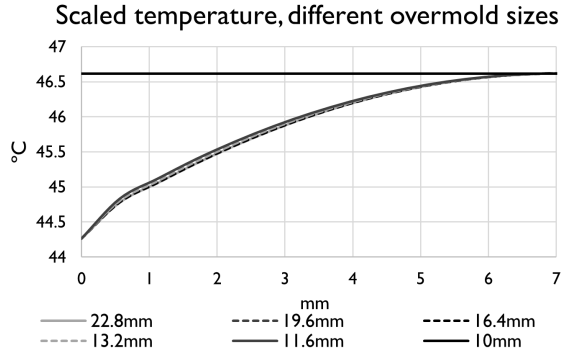


Figure 5.10: Diagonal of the temperature profiles at die level for different overmold sizes (in mm): curves are scaled to have the same maximum and minimum. The die stack is $10 \times 10\ mm^2$.

5.3.4 Simplified FEM

Based on the previous results, it seems that the complete removal of the FEM step in the correction methodology is difficult to be obtained. However, what can be exploited to simplify and to speed up the procedure is the knowledge that the curvatures of the temperature profiles, obtained for uniform power dissipation, are independent of the package structure, once these profiles have been scaled between the same minimum and maximum values. It is, indeed, sufficient to run a coarse and easy FEM to achieve an estimation of these two values for each specific package geometry and, then, scale a generic *basic surface*, which has been previously obtained for a generic package using a finer mesh, according to them.

The idea is, therefore, to consider a sort of *multi-scale approach*. In a first stage, the temperature responses of the stack configuration for the non-uniform and case-specific PMs are computed by the stack FTM. Then, on top of these results, the thermal impact of the spreading resistance is applied. This second step can be located on a lower level of accuracy, meaning that the inclusion of fine details is unnecessary and that the requirement for high accuracy can be relaxed. For these reasons, the package temperature response to uniform power dissipation can be computed by scaling a *basic surface* according to the case-dependent maximum and minimum values. These extreme values can be computed by implementing a FEM model with a coarse mesh, a quarter symmetry and further simplifications. In this way, the complexity and the computational effort required to compute $\Theta_{pack,unif}$ are reduced. In the next paragraphs, these simplifications and the algorithm itself are explained.

Scaling of a basic surface

In the previous Section, the similarity between $\Theta_{pack,unif}$ obtained for different structures, once they are scaled between the same extreme values, has been shown for overmolded packages with different overmold areas. However, the same basic surface can be considered also for different kinds of packages. This has been tested for various package configurations, such as low power and high power (as defined in Figure 5.4), different dimensions of PCB and overmold, extra copper layer on top, different materials and different BCs. In all these cases, all the scaled surfaces resulted to be comparable. It is possible that, for a completely different kind of package, the basic surface is significantly different. In this case, the extraction of $\Theta_{pack,unif}$ from a more detailed model is advisable. However, if small changes are performed afterwards in the development of the device, the new surface related to the modified package does not have to be recomputed and just an appropriate coarser model has to be run to obtain the correct maximum and minimum values.

Temperature computation level

According to the stack FTM methodology, the correction profile should depend on the level on which power is dissipated and on which temperature is computed. If N_p is the number of active layers and N_t the number of layers where the temperature has to be computed, this requires to run N_p FEM simulations for the package configuration and to extract $N_p \cdot N_t$ temperature profiles. However, as shown in Figure 5.11 (a), even if there is a difference in the obtained temperature values for profiles extracted at different levels, the corresponding correction profiles are comparable. The small error (<0.5%) close to the edges is negligible. This means that the basic surface can be obtained dissipating power on just one of the active layers.

Modeling of the stack in the package configuration

As mentioned before, since, from a multi-scale point of view, the correction step in the algorithm is located on a lower level of accuracy, some details can be neglected while computing $\Theta_{pack,unif}$. One of them is represented by the layered structure of the die stack. The die stack is, indeed, constituted by the stacking of different homogenized layers with different material properties (Si, μbumps + underfill and BEOL). However, many of these layers are much thinner compared to the others. The individual inclusion of the geometric characteristics and of the material properties of each of them enhances the complexity of the FEM simulation. Moreover, the information about the real layered structure is already included in the stack model and, therefore, in the uncorrected results obtained by the stack FTM. For these reasons, and also for the observation that the dependency of the correction profile on the power dissipation level is negligible, the die stack can

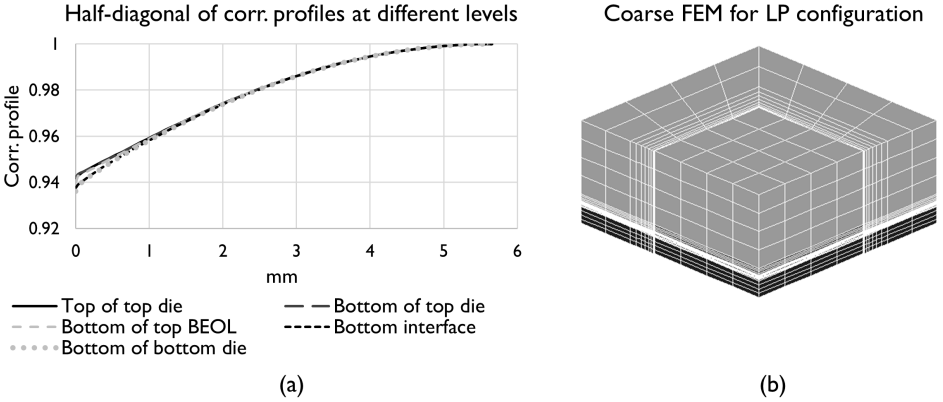


Figure 5.11: (a): Steady state correction profiles extracted at different levels. The LP configuration is considered here. More information about the dimensions are reported in Table 7.1. (b) Geometry of the coarse model used to compute $\Theta_{pack,unif}$.

be assumed to be of *one homogeneous material* while computing the maximum and minimum values of $\Theta_{pack,unif}$.

It is important, however, that $\Theta_{pack,unif}$ and $\Theta_{stack,unif}$ refer to the same level and to the same configuration. The purpose of their ratios, the correction profiles, is to account for the package spreading effect due to external parts not included in the FTM geometry. The thermal impact of the internal part, which is modeled both in the stack and in the package configuration, should not be included in the correction profiles. Different strategies can be implemented to make $\Theta_{pack,unif}$ and $\Theta_{stack,unif}$ comparable.

- Compute $\Theta_{stack,unif}$, on the same level where the maximum and minimum value of $\Theta_{pack,unif}$ are extracted in the FEM, making use of a lumped resistance approach. In this case, the die stack is assumed to be composed of full silicon in both the stack and the package configuration.
- Compute for the stack configuration, making use of appropriate resistance networks, the difference φ in temperature increase, on the level of interest, between the case of a full silicon stack and of a more realistic, layered die stack. The value of $\Theta_{pack,unif}$, computed considering a die stack made of full silicon, can, then, be translated according to φ to account for the neglected layered geometry. This approach is slightly more accurate than the previous one because it accounts for the *correct* value of the temperature on the level of interest.
- Compute the equivalent orthotropic thermal conductivity of the die stack and run the coarse model considering these values as the material properties of the whole stack. This approach is less straightforward but it's the only

one applicable for transient simulations. More details are reported in Section 5.4.3 and on the left hand side of Figure 5.18.

For the steady state configuration, the second option has been implemented.

Coarse FEM

The coarse model, from which the minimum and maximum temperature values are extracted, can, then, be created including all these considerations. It accounts for the simplified geometry, the dimensions, the BCs and the material parameters proper for each specific case. For the low power package shown in Figure 5.4, for example, a coarse FEM with a 1/4 symmetry, as the one illustrated in Figure 5.11 (b) has been used ($\approx$ 1900 elements). In case of a square geometry, a 1/8 symmetry can also be used, but the 1/4 symmetry is more general and can be applied for rectangular configurations (cf. Section 8.2). To have an accurate approximation of the minimum temperature at the dissipation level, a fine mesh is needed on the mold around the die and/or on the PCB (cf. Appendix A.2). Pillars and solder balls are not modeled individually but as uniform layers with equivalent material properties and, as explained in the previous Paragraph, the whole stack (dies + BEOL + interface layer) is assumed to be made of silicon.

Boundary conditions

Another issue comes from the definition of the BCs to be applied to the models of the stack and of the package configurations. Since they are both simplification of the same *real situation*, their results have to be related and the BCs selected in a proper way. One option is to choose the BCs to be applied on top and bottom of the stack configuration so that the maximum temperature increase for uniform power dissipation coincides with the one obtained in the package configuration for the same power dissipation scenario. In this way, the value of the correction profile in the center is approximately 1. However, since in the coarse model the stack is assumed of full silicon, when the equivalent heat transfer coefficients to be applied on the stack configuration have to be computed, the die stack in the stack configuration has also to be considered of full silicon.

Since for uniform power dissipation the heat path in the stack configuration is one dimensional (two directions), a lumped resistance network can be used to estimate proper values of the heat transfer coefficients. More precisely, h_t and h_b for the stack configuration can be computed by dissipating power on two different levels of the coarse package model, by storing the maximum obtained temperatures ($M1$ and $M2$) and by making use of the two corresponding resistance networks.

$$M1 = \frac{Q}{\frac{1}{R_{1,t}+1/h_tA} + \frac{1}{R_{1,b}+1/h_bA}} \qquad M2 = \frac{Q}{\frac{1}{R_{2,t}+1/h_tA} + \frac{1}{R_{2,b}+1/h_bA}}$$

where A is the floorplan area and $R_{1,t}, R_{1,b}, R_{2,t}$ and $R_{2,b}$ are, respectively, the conductive thermal resistance from the dissipation level to the top and to the bottom side of the stack configuration in case of power dissipated on level 1 and 2. These values are different because of the difference in dissipation levels.

Temperature for stack configuration

The other ingredient that is needed to compute the correction profiles is $\Theta_{stack,unif}$ for the layered structure. This value can be computed in two different ways. The former one is via the annulus method (cf. Section 3.4.1), taking advantage of the HSRs that have already been computed for the stack FTM. The latter one, instead, considers the resistance network for the stack configuration in which the real layered structure of the die stack is taken into account.

5.3.5 Flowchart of the steady state FTM algorithm

In order to allow an easy re-implementation of the developed FTM, the steps needed to obtain the temperature profiles for a packaged 3D-ICs are reported in the flowchart in Figure 5.12. Gray rectangles refer to computations preformed by FEM (Msc Marc [69]), white rectangles to computations performed in Matlab [66], while rectangles with rounded corners are used to indicate input and output quantities. The chart illustrates, in particular, how the correction profiles are obtained and how the temperature estimations, computed for a stack configuration, can be corrected in order to include the package thermal impact in steady state.

5.3.6 Results

The model setup used to validate the FTM including the package thermal impact is shown in Figure 5.13 for both the LP and the HP package configurations considered in this Chapter. Just 1/4 of the two structures is illustrated; both the FEM and the FTM have, however, been run for the complete geometry. The parameters concerning the geometry and the material properties are reported in Table 7.1, the only difference is that, for the cases considered in this Chapter, the dimensions of *Substrate1* are $13.6 \times 13.6 \times 0.5 mm^3$ and *Substrate2* does not exist. Note that, for the HP case, the package is not overmolded and the convective BC is applied on top of the die stack to mimic the effect of the heat sink (insulation is assumed on all other surfaces except for the bottom side of the substrate). The values of the heat transfer coefficients are reported in Table 5.1.

Figure 5.14 shows the results obtained by applying this methodology to these two package configurations. A non uniform power map with a hot spot in the center of the top die and four in the corners of the bottom die is considered (left column

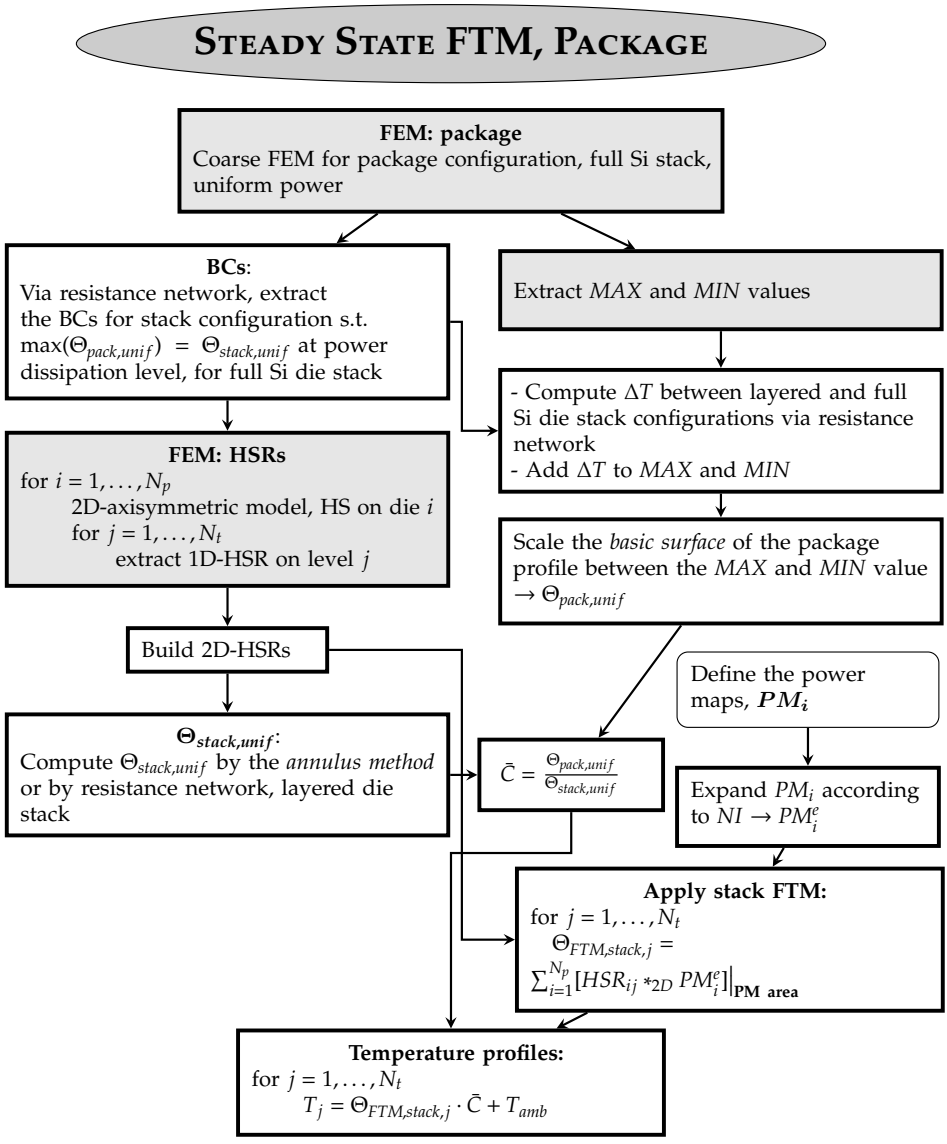


Figure 5.12: Flowchart representing the algorithm implemented for the steady state fast thermal modeling of packaged 3D-ICs.

of Figure 5.14). The dissipated power in the active regions is 1 W/mm^2 in the low power scenario and 5 W/mm^2 in the high power one. In the two graphs in the central column of Figure 5.14, where the results concerning the diagonal of the top and the bottom die in the low power configuration are reported, respectively, on the

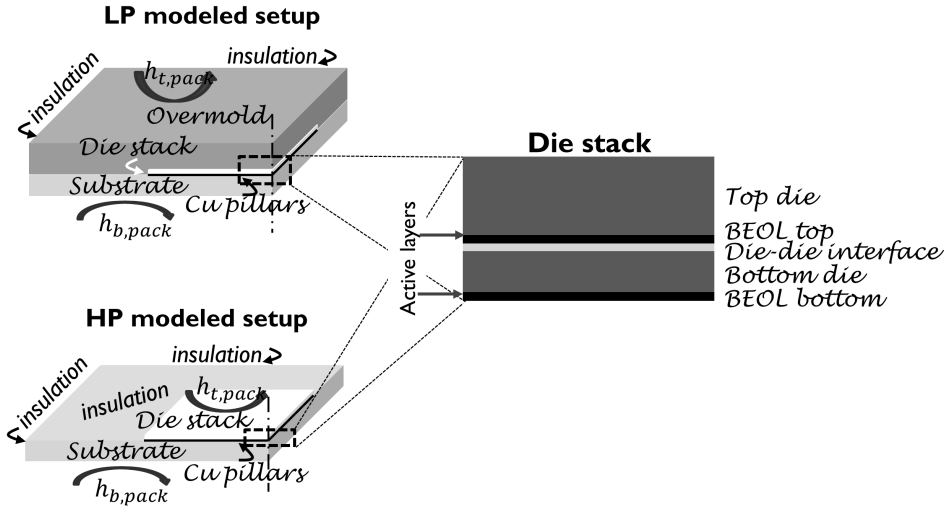


Figure 5.13: FEM setup used to validate the FTM including the package thermal effect (Figures 5.14 and 5.22). Just 1/4 of the geometry is illustrated.

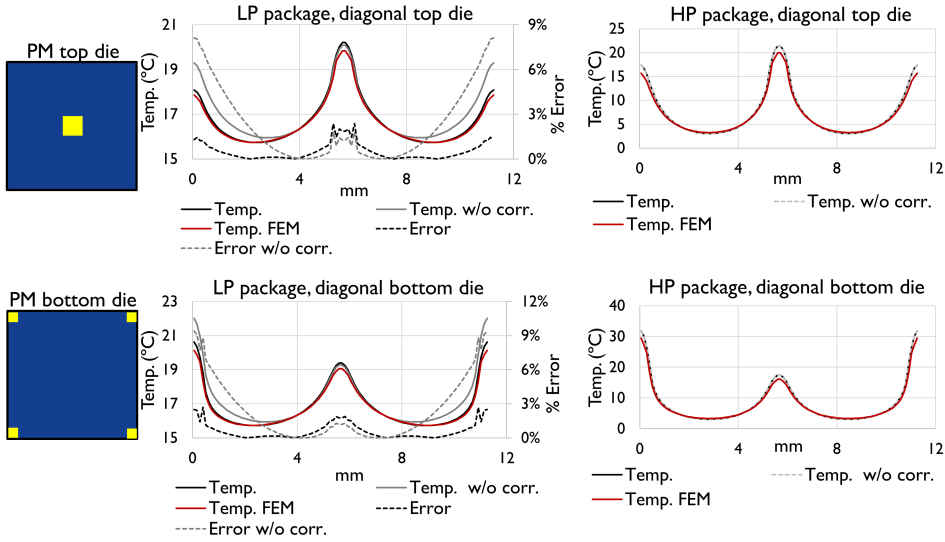


Figure 5.14: Diagonal of the temperature profiles obtained in the low power (central column) and high power (right column) configurations for the power maps illustrated on the left hand side. The FEM results are reported in red and the graphs concerning the low power configuration also report the percentage relative error with respect to them.

Low power		High power	
Boundary	Value	Boundary	Value
Top, $h_{t,pack}$	$0.00033 W/mm^2K$	Top of the stack, $h_{t,pack}$	$0.0146 W/mm^2K$
Bottom, $h_{b,pack}$	$0.00017 W/mm^2K$	Bottom, $h_{b,pack}$	$0.00017 W/mm^2K$
Lateral	insulation	Other boundaries	insulation

Table 5.1: Values of the heat transfer coefficients for the LP and the HP configurations used, both in the FEM and in the FTM models, to validate the package thermal inclusion in the FTM.

top and bottom row, the clear improvement in using this correction methodology is visible. The temperature curves (full lines) refer the left axis while the % relative error (dashed lines) to the right one. The red full line is used for FEM results and refers to the left axis. The error, especially in the corners, reduces from 8.1% to 1.3% by applying the correction procedure. The effect in the center is less pronounced because the BCs for the stack configuration have been chosen so that the stack model resembles the behavior of the package structure in this position. Moreover, the spreading *outside* the stack is more evident in the corners due to the presence of the larger overmold material and, as a consequence, the stack FTM (without correction) over-predicts the temperature in these locations.

Concerning the two graphs referring to the HP configuration on the right column, the effect of the package correction is almost inexistent. This is due to the high convective cooling rate from the top of the stack, which makes the upward heat path strongly predominant compared to the lateral spreading. On top of the high convection, which quickly removes the heat, the lack of a spreading structure on top of the stack results in a minimum spreading effect. This is why the corrected and the uncorrected temperature profiles are the same and they are very similar to the FEM one.

These two cases show that a package correction is not always needed, but its need is highly case dependent. However, even if the correction is applied in an unneeded situation, as in the high power case, the accuracy of the method does not deteriorate. A possible metric to estimate the significance of a package correction, i.e. the heat spreading due to the package, can be related to the maximum and minimum temperature values in the profile extracted from the coarse FEM (the one that is used to build the correction profile). More precisely, the spreading metric can be defined as

$$sp = \frac{MAX - MIN}{MAX}. \tag{5.10}$$

This metric is, in particular, directly related to the improvement achievable in temperature estimation in case of uniform power dissipation. For non-uniform PMs, the *sp* value represents an upper bound for the improvement that can be actually achieved by applying the package correction algorithm. How much this

upper-bound deviates from the real value is highly dependent on the dissipated PM. For two considered configurations, for example, $sp = 6.85\%$ for the low power scenario and $sp = 0.002\%$ for the high power one. The difference in their magnitude confirms the uselessness of the package correction in the second case. It is important to note that the added computational costs associated to this methodology, with respect to the one for the stack configuration, consists in the time needed to built and run the coarse FEM (running time less than 1 sec).

5.4 Transient regime

In this Section the package correction strategy, presented in Section 5.3 for the steady state, is extended to the transient regime. The main difference with respect to the steady state methodology is that the time dependency of the heating process has to be taken into account. In order to include the package thermal impact in an appropriate way, the basic algorithm of the transient FTM for the stack configuration has to be revised as explained in Subsection 5.4.1. The new transient procedure, which includes the package correction, allows to achieve a much better accuracy but, at the same time, causes an increase in computational time. These results have been published in [61].

5.4.1 Methodology

As explained in Section 2.5.3, for a stack configuration the time dependent temperature profile on level j due to power dissipated on level i can be computed in two different ways: 3D-convolution or 2D-convolution with subsequent time superposition. The related numerical formulas are

$$\Theta_{z_i}(\cdot,\cdot,z_j,\bar{t}_k) \approx \sum_{\bar{t}_l=1}^{\bar{t}_k} \bar{\Theta}_{z_i}(\cdot,\cdot,z_j,\bar{t}_k;\bar{t}_l) = \sum_{\bar{t}_l=1}^{\bar{t}_k} HSR_{z_i}(\cdot,\cdot,z_j,\bar{t}_l) *_{2D} PM_i(\cdot,\cdot,\bar{t}_k-\bar{t}_l), \quad \forall \bar{t}_k, \tag{5.11}$$

$$\Theta_{z_i}(\cdot,\cdot,z_j,\cdot) = HSR_{z_i}(\cdot,\cdot,z_j,\cdot) *_{3D} PM_{z_i}. \tag{5.12}$$

It is worth to repeat here that the main difference between the 3D- and the 2D-approach is that, in equation (5.12), the solution is computed, at the same time and by means of one single operation, for all the possible values of both the spatial and the temporal variables. However, for this approach, additional time layers, constituted by zero's matrices, have to be added to the HSRs in order to locate data referring to the present time in the middle of the time vector. When the approach in equation (5.11) is selected, on the other hand, the results refer to a fixed point in time t_k ($t_k = \bar{t}_k \Delta t$) and, even to obtain $\Theta_{z_i}(\cdot,\cdot,z_j,t_k)$ at fixed

time t_k, multiple 2D-convolutions operations are needed, one for each past power dissipation time, $t_k - t_l$. Due to the much larger number of operations needed to perform the approach in equation (5.11), 3D-convolution normally remains computationally much faster than 2D-convolution plus time superposition (cf. Section 2.5.3), despite the larger HSRs (extra zeros) used in equation (5.12).

The temporal sequence of the PMs has also to be considered for the application of the package correction procedure. If the same power $\tilde{q}$ is dissipated at time $t_k - t_0$ rather than at $t_k - t_1$, with $t_0 \neq t_1$, the system response at time t_k is different. For this reason, each impulsive partial temperature increase profile $\bar{\Theta}_{z_i}(\cdot, \cdot, z_j, t_k; t_l)$ needs to be corrected individually, depending on the value of $t_k - t_l$. In this way, the different time constants of the different parts of the package are taken into account. This is why the transient FTM with package correction is implemented via 2D-convolution and time superposition. The 3D-convolution algorithm does not allow, indeed, direct access to the partial temperature increase profiles.

Since the HSRs are computed for impulsive heat dissipation and, therefore, $\bar{\Theta}_{z_i}(\cdot, \cdot, z_j, t_k; t_l)$ are the partial temperature profiles at time t_k due to impulsive power dissipation at time $t_k - t_l$, the transient correction profiles for the inclusion of the package thermal impact need to be computed accordingly, taking time dependency into account. This means that the system responses for the package and the stack configurations are computed for uniformly distributed, impulsive power dissipation and they are stored as functions of time, until steady state is reached. Analogously to the steady state methodology, the time dependent correction profiles are computed as

$$\bar{C}_l = \frac{\bar{\Theta}^l_{pack,unif}}{\bar{\Theta}^l_{stack,unif}} \tag{5.13}$$

where $\bar{\Theta}^l_{pack,unif}$ and $\bar{\Theta}^l_{stack,unif}$ are the temperature profiles obtained, respectively, for the package and the stack configurations at time step $\bar{t}_l$ in case of uniform, impulsive power dissipation during the first time step, i.e. during the time interval $[t_0, t_1)$.

In Figure 5.15, the half-diagonals of the correction profiles are plotted for the two package configurations considered in this Chapter at different time steps. The full lines refer to correction profiles computed at the end of the power pulse (50 msec, since $\Delta t = 0.05$ sec), the dashed curves to the correction profiles after 0.3 sec (0.05 sec power pulse plus 0.25 sec cooling) while the dotted curves to profiles obtained after 1 sec from the beginning of the simulation. The dependency of the correction factors $\bar{C}_l$ on time is clearly visible. To be noted from the same plots is the apparent lack of normalization of the profiles. This is because the BCs applied to the stack configuration have been selected so that, for uniform and continuous power dissipation, the maximum temperature increases at steady state, for the stack and the package geometry, are approximately equal. While, in steady state, the application of the correction profiles compensates just for the difference in

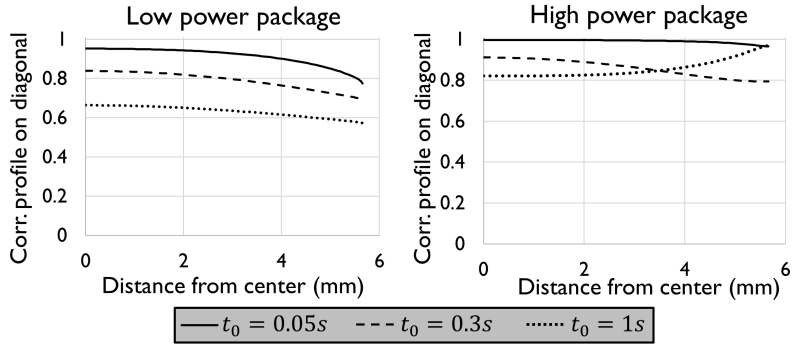


Figure 5.15: Half diagonal cross sections of the correction profiles extracted for the low power (left) and the high power (right) package configurations at different times: at the end of the HS dissipation at 0.05 sec (full line), at 0.3 sec (dashed lines) and at 1 sec (dotted lines).

thermal resistance experienced in different locations due to the package thermal spreading, in transient regime the capacitive capability of the different materials needs also to be taken into account. Since, in the stack configuration, parts of the package are neglected, the correction profiles need to account for their missing capacitive effect. This is why the maximum value of $\bar{C}_l$ is not 1 and why it is time dependent.

In Figure 5.16, the correction procedure for the calculation of the temperature profile due to power dissipation in a specific location $\boldsymbol{x}_a = (x_a, y_a)$ is illustrated. For clarifying reasons, in this example the impact of the power dissipated in neighboring locations is neglected, i.e. time varying power dissipation is assumed just in one location in the die and everywhere else it is set to zero. For more general situations, the PM-HSR multiplications reported in the graph should be substituted by 2D-convolutions. The three plots on the top of the Figure show the applied PM while the ones in the middle the HSR. The illustrated procedure shows how to compute the temperature at $t_k = 0.3$ sec with a time discretization of $\Delta t = 0.05$ sec (i.e. $\bar{t}_k = 6$). The first step consists in computing, for each value of $t_l \leq t_k$, the impulsive partial temperature increases $\bar{\Theta}(\boldsymbol{x}_a, t_k; t_l)$ of the system. These originate from power dissipated in the past, at $t_k - t_l$, and they take into account how long ago (t_l) each pulse has been generated. Secondly, each single obtained $\bar{\Theta}(\boldsymbol{x}_a, t_k; t_l)$ is corrected by means of multiplication with the correction profile's value corresponding to its location in space and to t_l, the amount of time passed from dissipation (third row in Figure 5.16). These operations are normally point-by-point multiplications between 2D-matrices but, in this setting, they reduce to single multiplications since the spatial dimension is not considered. Finally, exploiting superposition principle, the corrected impulsive partial temperature increases are summed up to provide the temperature increase profile at time t_k, which account for both the power dissipation history and the package thermal

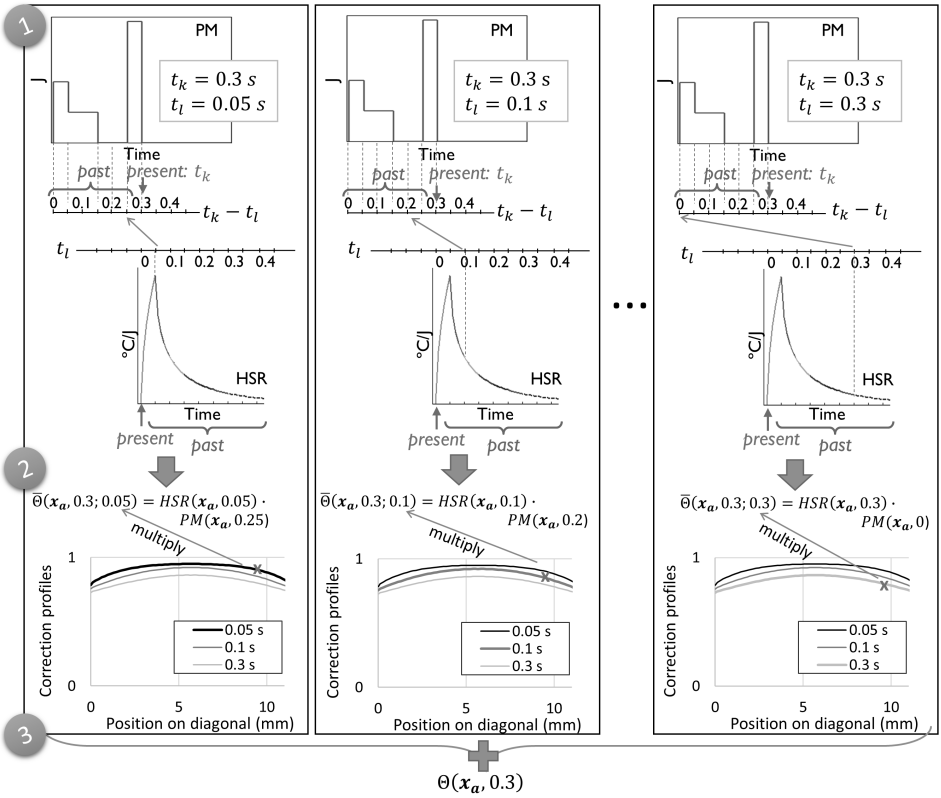


Figure 5.16: Schematic of the transient FTM methodology with package correction. This illustration assumes that power is dissipated in only one point, $\boldsymbol{x}_a$, and that the temperature is computed just in $\boldsymbol{x}_a$. The first row of the Figure shows the applied PM, the second row the HSR and the third row the correction profiles.

impact.

5.4.2 Computational time analysis

As already explained in Section 2.5.3, the application of 2D-convolution and time superposition instead of 3D-convolution has an impact on the overall computational time. This impact depends on the spatial resolution of the HSRs and of the PMs as well as on the simulated time and the time constant of the system. Concerning the package correction procedure, once the correction profiles are computed (cf. Section 5.4.3), the implementation of the package corrected FTM consists in their point-by-point multiplication with the intermediate 2D-

convolution impulsive partial temperature increase results, $\bar{\Theta}$. The computational cost of these operations can be neglected when related to the convolution one.

For this reason, the comparison in computational time between the package corrected FTM and the stack FTM can be performed comparing 2D- and 3D-convolution approaches. This comparison has already been reported in Section 2.5.3. For that case study, for example, 30 simulated time steps result in a 5 times slower algorithm when the package correction is applied. However, despite this slowdown with respect to the stack FTM, once the package FTM is compared to analogous FEM, it provides really accurate results in much shorter computational time. For the case presented in Section 5.4.5, consisting in a $100\mu m \times 100\mu m$ spatial grid and 32 time steps, for example, the FEM for a low power configuration is 170 times slower than the corresponding package, 2D-convolution based FTM (20 hours versus 7 minutes simulating time).

5.4.3 Correction profiles

As for the steady state regime, the application of the correction profiles can be seen as a multi-scale strategy, in which the correction is located on a lower level of accuracy. This means that the FEM used to compute $\bar{\Theta}^l_{pack,unif}$ can be coarsened and, by doing so, computational time can be saved. Different steps, which are not only related to the coarsening and simplification of the FEM, have been implemented in order to reduce the computational time needed to obtain the time dependent correction profiles. These steps are illustrated in the following of this Section.

- Paragraph *"Temperature for uniform power dissipation in the stack configuration"* explains how the time dependent temperature increase due to uniform power dissipation in a stack configuration, $\bar{\Theta}^l_{stack,unif}$, can be efficiently computed. These values are used to compute the correction profiles $\bar{C}_l$ as shown in equation (5.13).
- In Paragraph *"Levels of the correction profiles"* the sufficiency of one correction profile per time step, independent of the level where power is dissipated, is shown.
- In Paragraph *"Equivalent material properties"* the possibility to use a single material to model the die stack in the package configuration is presented. In this way, the FEM is simplified because the layered structure of the die stack is neglected. However, equivalent material properties should be used to properly replace this layered structure with a single material.
- In Paragraph *"Temperature profile extraction"* an efficient way to extract $\bar{\Theta}^l_{pack,unif}$ from the results obtained by the coarse FEM is explained.

- In Paragraph *"Error metric"* an error metric is presented to estimate in advance the improvement achievable by applying the package correction strategy on top of the transient stack FTM. This metric is proposed for uniform and continuous power dissipation. It can, nevertheless, give an idea of if it is worth to use the computationally more expensive algorithm including the package thermal impact or not.

Temperature for uniform power dissipation in the stack configuration

In case of uniform power dissipation on any horizontal layer of a stack configuration, the system temperature response on each level is constant in space but variable in time. This means that, for each given power dissipation and temperature response level, one single value per time step is enough to describe the thermal behavior of the system. These values, $\bar{\Theta}^{l}_{stack,unif}$, which are needed to build the correction profiles, can be easily computed by the FTM via 3D-convolution between uniform PMs and HSRs. However, since the resolution of the HSRs is the one required for the final temperature profiles, the direct application of the stack FTM results in *highly resolved*, *uniform* temperature fields. Since a single value per time step is enough to fully characterize the temperature response, this is unnecessary and time consuming.

For this reason, the *annulus method* has been applied in transient regime (cf. Sections 3.4.1 and 3.6). The methodology, originally presented for steady state, is applied to transient regime by just considering each time step separately. This returns a sequence $\bar{\Theta}^{l}_{stack,unif}$ of numbers representing the temperature increase at time step $\bar{t}_l$ due to uniform power dissipation in the time interval $[t_0, t_1)$.

Levels of the correction profiles

According to the convolution based FTM methodology, the correction profiles should depend on the levels on which power is dissipated and on which temperature is computed. However, as for the steady state regime, this condition can be neglected while extracting the correction profiles. Figure 5.17 demonstrates this claim. The picture shows the half-diagonal cross sections of the correction profiles at different times (0.05 sec in black, 0.3 sec in dark gray and 1 sec in light gray), for the two different kinds of package considered in this Chapter, the low power (left) and the high power (right) configurations. Different marker's types indicate different dissipation and temperature response levels for which the correction profiles are computed at the specific time. In both analyzed cases, neglecting the information about the power dissipation and temperature computation levels doesn't cause relevant loss in accuracy while saving computational time. For this reason, in the following, a single, time dependent, correction profile is extracted for each package.

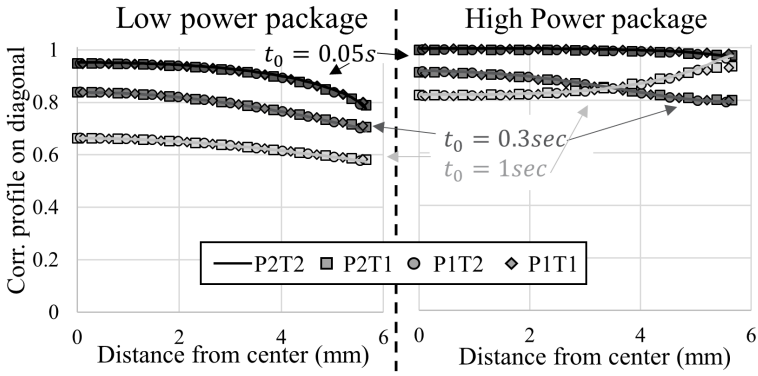


Figure 5.17: Half diagonal cross sections of the correction profiles extracted for the low (left) and high (right) power package configurations at different times: at the end of the HS dissipation at 0.05 sec (black), at 0.3 sec (dark gray) and at 1 sec (light gray). Different marker's types indicate different dissipation and temperature response levels in which the correction profiles are computed. In the legend *PxTy* stands for power dissipated on die *x* and temperature computed on die *y*; 1 indicates top die and 2 bottom die.

Equivalent material properties

For the same reasons presented in Paragraph *"Modeling of the stack in the package configuration"* in Section 5.3.4 for the steady state regime, the die stack in the package configuration can be considered made of just one material also in the transient regime while computing $\bar{\Theta}^l_{pack,unif}$. However, the approach based on resistance networks, which has been implemented for the steady state, cannot be applied in the transient regime. The aim of that step in the steady state algorithm was to retrieve the correct temperature increase for a layered stack in the package configuration even if a single material is considered for the stack in the modeled package geometry. The reason why this is not possible is that, in the transient regime, also the capacitive effect of each layer has to be taken into account. This means that capacitances have to be added to the network. However, the construction of an RC-network for the transient regime is not as straightforward as the construction of a resistance network for the steady state. For this reason, an approach based on the calculation of equivalent material properties for the stack, including the resistive and capacitive effects of all the materials, have been considered. The purpose of this *equivalent*, *uniform* material block is to mimic the thermal response of the *layered* die stack. The procedure adopted to compute its thermal properties is explained hereafter.

For each specific configuration, the equivalent orthotropic thermal conductivity is computed, at steady state, by means of resistance equivalent networks for the stack

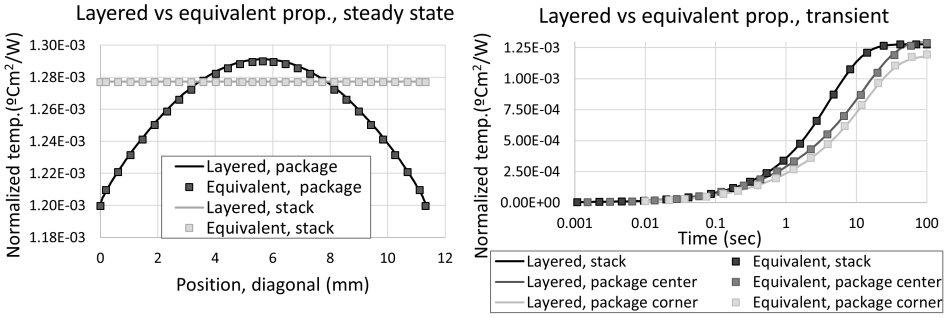


Figure 5.18: Normalized FEM temperatures for uniform power dissipation obtained by using equivalent properties (markers) and by using the more realistic layered structure for the die stack (full line). Left: diagonal cross sections in steady state regime for the package (black) and stack (gray) configurations. Right: logarithmic time scale evolution in transient regime for the stack (black) and the package configuration, in the center (dark gray) and in the corner (light gray) of the die.

configuration. Concerning the out of plane value, k_z, the temperature increase, ΔT, due to uniform power dissipation on the temperature extraction level is computed, as a first step, for the layered stack configuration. Once ΔT is known, a reverse problem can be solved and the resistance needed to get the same ΔT, with a uniform material block representing the die stack, can be calculated. From this value, based on the geometric properties, the equivalent $\tilde{k}_z$ value can be retrieved.

Regarding the in-plane thermal conductivity, $\tilde{k}_{x,y}$, just the die stack section of the stack configuration is considered in the computations. The thermal resistance of each layer is computed for a horizontal heat path, i.e. $R_{th} = \frac{cs}{k_{x,y} \cdot cs \cdot l}$ where cs is the chip size, l and $k_{x,y}$ are, respectively, the thickness and the in-plane thermal conductivity of each specific layer. Their parallel connection provides an equivalent resistance from which the equivalent $\tilde{k}_{x,y}$ can be extracted.

The so obtained orthotropic conductivity values are assigned to the die stack equivalent material in the package configuration. This basically means that what is named "die stack" in Figure 5.4 is substituted by a block of homogeneous material while computing the package thermal response in the FEM model. On the left hand side of Figure 5.18, the very good agreements between the normalized temperature profiles obtained, on the diagonal of the die, by using equivalent properties and by using the more realistic layered structure for the die stack, are shown for both the stack and the package configuration. Concerning the "more realistic structure", this is modeled considering the individual layers (dies, interface, BEOL, . . .) in the die stack, as they are shown in the central section of Figure 5.4.

For simulations in transient regime, the equivalent capacitance value is also needed. Since the thermal capacitance, C, is defined as a volumetric integral $C = \int_V c\rho dV$

where c is the specific heat capacity, ρ the mass density and V the volume, its equivalent value is computed by means of volume average. The right hand side of Figure 5.18 shows the comparison between the normalized transient thermal responses of a system modeled using equivalent properties and using the more realistic layered structures. The plot shows the time dependency of the temperature responses, in logarithmic time scale, for the stack geometry (black) as well as for the package configuration. In this last case, since results are space dependent, two curves are shown, one referring to the center (dark gray) and the other one to the corner of the die (light gray). The two graphs in Figure 5.18 prove the possibility to substitute, in the package FEM simulations, the different layers in the die stack with a single material block, to which equivalent orthotropic properties are assigned. Computational time is, consequently, reduced.

Temperature profiles extraction

Another step in the algorithm, in which computational time can be saved, is the extraction of the temperature profiles from the package FEM model. The correction procedure consists in *point-by-point* multiplication between the impulsive partial temperature increases obtained by the stack FTM for a case dependent PM and the correction profiles. For this reason, the correction profiles and, in particular, $\bar{\Theta}^{l}_{pack,unif}$, which are extracted from the coarse model, need to have the same high resolution as the final temperature profiles. The common way to achieve this aim is through space interpolation of $\bar{\Theta}^{l}_{pack,unif}$ at each individual time step.

Another option, which allows to save computational time, is to take into account what each single correction profile is applied to. The overall temperature increase, $\Theta_{z_i}(\cdot,\cdot,z_j,t_k)$ in the FTM is, indeed, computed as the sum of impulsive partial temperature increases, $\bar{\Theta}_{z_i}(\cdot,\cdot,z_j,t_k;t_l)$, due to impulsive power dissipation in the past $(t_k - t_l)$. The further ago an impulse has been dissipated, the less its contribution on the final temperature increase is. For this reason, the required accuracy with which $\bar{\Theta}^{l}_{pack,unif}$ are extracted can be lowered according to how long ago the generating impulsive power has been dissipated.

After several tests on different package structures, it has been found that the difference in shape between the temperature profiles obtained at different time steps for uniform, impulsive, power dissipation in a package geometry, $\bar{\Theta}^{l}_{pack,unif}$, can, in most cases, be neglected. What does change, are the achieved maximum/minimum values of the temperature profiles. For this reason, the fine grid interpolation of just a couple of profiles, followed by a surface scaling in order to match the maximum/minimum values predicted at each specific time step by the coarse FEM, proved to be enough.

These two basic profiles, which are obtained by spline interpolation of the data extracted from the coarse FEM results, are the first two in chronological order.

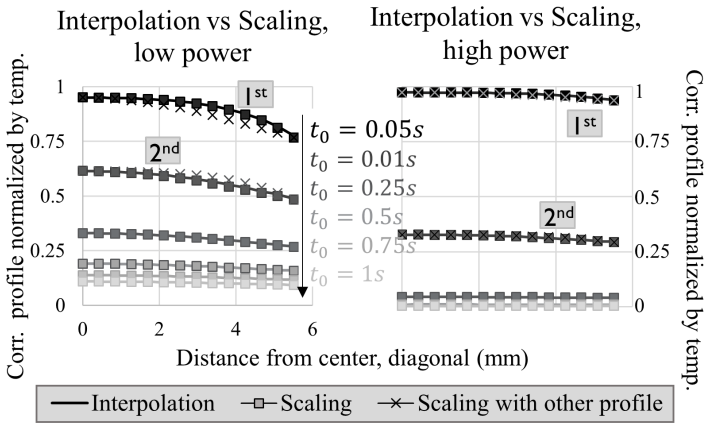


Figure 5.19: Normalized correction profiles at different time steps for the low power package (left) and the high power one (right). Full lines represent results obtained by interpolation at each single time step while square markers the ones got by the scaling approach. Crosses are for results obtained if just one correction profile, instead of two, would have been interpolated and then scaled.

These are the ones with the highest impact on the final results and, for which, therefore, higher accuracy is more appropriate. The creation of the highly resolved temperature profiles for all the other time steps is performed by scaling the second interpolated surface, the one obtained a time step after the cooling down phase has begun. Differently from the steady state scenario, therefore, the basic surfaces are case dependent but, once the cooling down phase has begun, they can be considered time independent. Since the scaling approach requires less interpolating steps than a full interpolation approach, it is computationally convenient. The gain is proportional to the amount m of time steps needed to reach steady state. For 20 time steps, for example, the speed up achieved in this step of the algorithm is around 10 times.

Normalized correction profiles at different time steps are shown in Figure 5.19 for the low power package, on the left, and for the high power package, on the right. The normalization is performed multiplying each real correction profile by the ratio between the maximum temperature obtained at that specific time and the one at the end of the impulsive power dissipation. This normalization is used to account for the quantity to which each correction profile is applied in the FTM algorithm. Since impulsive power dissipation is considered at each time step, the correction profiles related to later time steps are normally applied to temperature maps with lower values. Different colors in the graphs refer to different periods of time after the beginning of the impulsive power dissipation. Results obtained by interpolation at each single time step are indicated by full lines while square marks are used to represent the scaling approach. As we can see, the introduced

loss in accuracy is negligible.

The crosses in Figure 5.19 denote the results that would be obtained by interpolating just one temperature profile instead of two. The crosses referring to the normalized correction at $t = 0.05$ sec are obtained by scaling the temperature profile referring to $t = 0.1$ sec and vice versa. For the high power package, the interpolation of two separate temperature profiles may be avoided but this is not the case for the low power configuration. This is mainly due to the lower cooling rate and higher spreading resistance in the latter situation. However, for general situations, the interpolation of two temperature profiles proved to be better and, therefore, it will be used in the rest of the thesis.

Error metric

As already explained, the implementation of the package correction algorithm in transient regime requires much higher computational time than the transient FTM for the stack configuration. This means that, if for certain specific package structures and cooling solutions, the package thermal impact is low, the 3D-convolution based algorithm can be applied and the correction procedure avoided. For this reason, an a priori estimation of the maximum relative improvement (*impr*) achievable at time t_k, by applying the transient correction procedure on top of the stack FTM, has been derived for uniform power dissipation, continuous in time.

For $t_k = \bar{t}_k \Delta t$, let's define $\max(impr^k)$ as

$$\max(impr^k) = \max\left|\frac{\text{err reduction}^k}{\text{exact solution}^k}\right| = \max\left|\frac{(\Theta^k_{stack,unif} - \Theta^k_{exact,unif}) - (\Theta^k_{pack,unif} - \Theta^k_{exact,unif})}{\Theta^k_{exact,unif}}\right| \approx \frac{(\Theta^k_{stack,unif} - \min(\Theta^k_{pack,unif}))}{\min(\Theta^k_{pack,unif})} \tag{5.14}$$

where $\Theta^k_{stack,unif}$ and $\Theta^k_{pack,unif}$ are, respectively, the temperature profiles at time step $\bar{t}_k$ calculated for the stack and for the package configurations in case of uniform power dissipation, *continuous* in time. $\Theta^k_{exact,unif}$ is the corresponding exact solution for the real configuration.

Due to the definition of the BCs, which ensures that, at steady state, the temperature of the FTM for the stack configuration approximately matches the maximum one of the package configuration, the maximum error reduction is achieved in the corner of the die, where the spreading effect is higher. Moreover, being $\Theta^k_{exact,unif}$ unknown, it is approximated by $\Theta^k_{pack,unif}$ and the maximum of the ratio representing $impr^k$ is achieved for $\min(\Theta^k_{pack,unif})$, the value in the corner. This information can be easily obtained from the coarse FEM results for the package configuration. The minimum temperature data, *MIN*, needed in the scaling phase (cf. previous

Paragraph *"Temperature profile extraction"*), can be used to this aim. The only difference is that those data are obtained for impulsive power dissipation while the error metric is derived for continuous power. A cumulative sum of all these minimum values provides the desired quantity as function of time. $\Theta^k_{stack,unif}$ can be computed using the algorithm illustrated in Paragraph *"Temperature for uniform power dissipation in the stack configuration"* at the beginning of this Section. The cumulative sum of the values obtained by that algorithm provides the temperature increase for continuous power dissipation in the stack configuration starting from the corresponding temperature response to an impulsive power source. This results in the following formula:

$$\max(impr^k) = \frac{\text{cumsum}(\bar{\Theta}^k_{stack,unif}) - \text{cumsum}(MIN)}{\text{cumsum}(MIN)}. \tag{5.15}$$

Despite the definition of the BCs for the FTM, $\bar{\Theta}^k_{stack,unif}$ cannot be substituted by the cumulative sum of the maximum temperature values of the coarse FTM. This is because the match between the maximum temperature values in the stack and the package configuration is imposed for steady state. The significant role played by the difference in thermal capacitance between the two configurations needs to be taken into account when considering the relative improvement achievable from modeling a package rather than a stack configuration.

The estimation of the relative error reduction can be easily and quickly computed. Even if it is derived for uniform power dissipation, constant in time, it provides useful information about the thermal impact of the package. In this way, it is possible to decide a priori if the improvement, achievable including the package effect, justifies the higher computational time of the 2D-convolution based methodology. As for the steady state regime, max(*impr*) represents an upper bound to the improvement in accuracy achievable by applying the package correction: for realistic cases, with non-uniform and time varying power maps, the effective improvement can be much lower.

Figure 5.20 (a) reports the computed max(*impr*) as a function of time for the LP (black) and for the HP (gray) configurations up to 1.6 sec. As it was already clear from the steady state analysis, the package thermal impact is much more relevant in case of the LP package. Figure 5.20 (b) shows the temporal evolution, until steady state, of the quantities that appear in equation (5.15). The huge difference in thermal capacitance between the stack and the package geometry in the LP configuration, together with the high spreading resistance, is the reason why the relative maximum improvement increases and achieves a value higher than 0.5 at 1.6 sec. A significant time delay between the package and the stack configuration is, indeed, clearly visible from the right hand side graph. While approaching the steady state regime, the time delay reduces and the same will happen to max(*impr*). At steady state, indeed, the package correction accounts only for the difference in thermal spreading between the two models, not for the difference in capacitance.

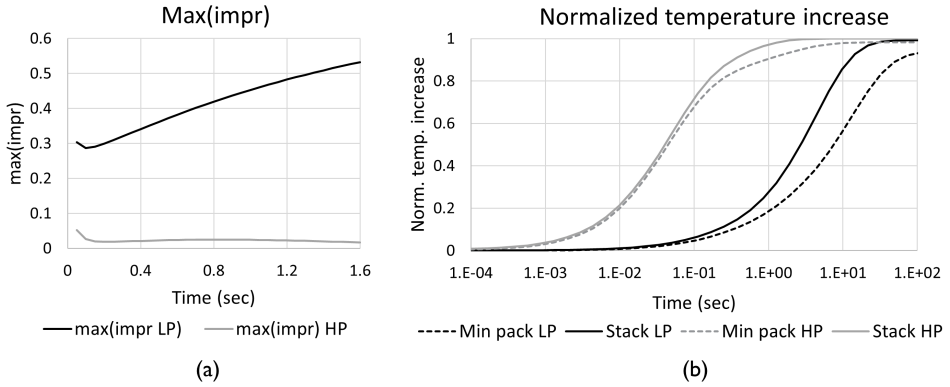


Figure 5.20: (a) Estimation of the maximum relative improvement in case of the LP and the HP packages. (b) Time evolution of the temperature response to uniform power dissipation in case of the HP and the LP packages. The curves refer to the stack configurations and to the corner of the die in the package configurations.

Concerning the HP package, the impact of the correction is much lower for two reasons: 1) there is less difference between the capacitance of the two models because there is less material in the package configuration that is not considered in the stack configuration and 2) the lateral spreading resistance is lower than in case of the LP package due to the good cooling solution applied on top of the device.

5.4.4 Flowchart of the transient FTM algorithm

In this Subsection the steps needed to include the package thermal impact in the transient FTM are illustrated in the flowchart in Figure 5.21. Since the increase in computational time, due to the implementation of the package correction algorithm, is not always followed by a significant improvement in accuracy, a decision block (diamond), based on the developed error metric, appears in the flowchart. The main steps of the algorithm are also explained hereafter.

1. Extraction of the BCs, to be applied to the stack configuration, from the FEM steady state coarse model for the package configuration in case of uniform power dissipation. The BCs are defined in such a way that the maximum temperature in the package configuration approximately matches the one in the stack configuration. In this step, the uniform die stack material is considered to be silicon in both the package and the stack configuration. This can create some inaccuracy in the following steps, when layered die stacks are considered. However, these errors are mainly canceled out by the application of the correction profiles (cf. Paragraph *"Boundary conditions"* in Section 5.3.4).

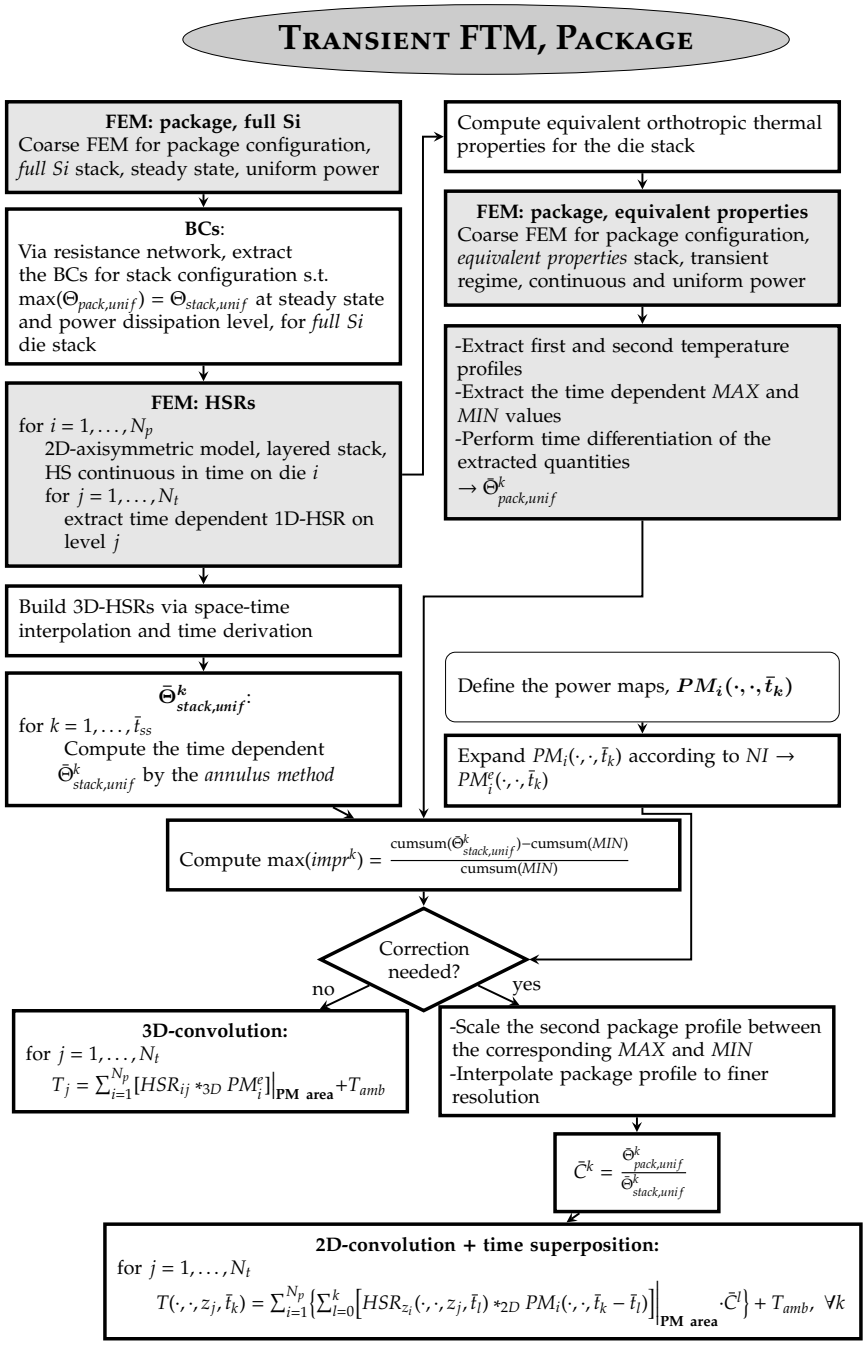


Figure 5.21: Flowchart representing the algorithm implemented for the transient fast thermal modeling of packaged 3D-ICs.

2. FEM computation of transient HSRs: 2D-axisymmetric models for a stack configuration with layered die stack and hot spot dissipation. The power dissipation is assumed to be constant in time. Time differentiation is performed afterwards to account for impulsive power (cf. Paragraph *"Transient"* in Section 2.5.3).

3. Calculation of the equivalent die stack properties (cf. Paragraph *"Equivalent material properties"*).

4. Extraction of the time dependent package temperature profiles from the FEM transient coarse model, with equivalent die stack properties and uniform, continuous power dissipation. The model is run until steady state is reached. The full package temperature profiles are extracted for the first two time steps, while, for all the others, just the maximum and the minimum values are stored. Time differentiation is performed to account for impulsive power.

5. Calculation of $\bar{\Theta}^k_{pack,unif}$ through spline interpolation of the two fully extracted temperature profiles in order to have the same high resolution as in the stack FTM. Scaling of the latter profile according to the extracted extreme values for all the other time steps (cf. Paragraph *"Temperature profile extraction"*).

6. Computation of $\bar{\Theta}^k_{stack,unif}$, the uniform temperature profiles in the stack configuration, using the fast algorithm presented in Paragraph *"Temperature for uniform power dissipation in the stack configuration"*.

7. Computation of the relative improvement estimation (cf. Paragraph *"Error metric"*).

If, from the calculation of the relative improvement, a correction is needed:

8 a. Computation of the correction profiles (cf. Section 5.4.1).

9 a. Implementation of the 2D-convolution plus package correction algorithm to compute the final temperature profiles (cf. Section 5.4.1).

Otherwise

8 b. Implementation of the 3D-convolution for the transient FTM without package correction (cf. Section 2.5.3).

5.4.5 Results

In order to prove the accuracy of the transient FTM methodology including the package impact, comparisons have been performed with respect to validated

FEM models of the complete package structure, not just of the die stack as in Chapter 3. The case shown in the following of this Section refers to a face-to-back stack of two 8 $mm \times 8$ mm dies. The stack is considered to be packaged in a low power configuration as the one shown in Figure 5.13. The same geometry, material and BCs parameters used for the steady state validation are considered also in this case (cf. Tables 5.1 and 7.1; the dimensions of *Substrate1* are, however, $13.6 \times 13.6 \times 0.5 mm^3$ while *Substrate2* does not exist). A resolution of 100 $\mu m \times 100$ μm is assumed in space while a step of 50 msec is considered in time. The power dissipation is non-uniform in space and non-constant in time. The simulated time is 1.6 sec, which is shorter than the time constant of the system. For this reason, also the HSRs are recorded until 1.6 seconds.

FEM validation

Figure 5.22 (a) shows the maximum predicted temperature increase on the top and the bottom die as a function of time. Since the power map varies with time, the location of the maximum temperature is not fixed. Blue color refers to the bottom die while red to the top die. Full lines represent the results obtained by the corrected FTM, dashed lines by the stack FTM and circles the ones from the FEM, with respect to which the FTM is validated. As it is visible, a significant improvement is achieved by applying the correction procedure.

For this test case, the selection of the more computationally expensive algorithm allows to keep the error at maximum temperature always below 3°C (5%) (Figure 5.22 (b)). The choice of the faster FTM option would results in an error up to 11°C (35%) at the end of the simulation. It is worth to note that the error referring to the stack FTM methodology has a tendency to grow with time. This is due to the increasing impact that the package assumes during chip activity (cf. Figure 5.20 (a)). Immediately after power dissipation, just the die stack and a small part of the package affect the temperature rise. With the passing of time, more sections of the package enter the heat path having, therefore, an impact on the final result. Including the package correction in the algorithm allows to take all these aspects into account. In this case, indeed, the error doesn't show any particular trend but it is maintained at really low values.

The results shown up to here refer to the maximum temperature location at each time step, which, in this case, never happens to be in the corners. This means that the accuracy improvement could be higher in other locations, where the effect of the correction profiles is higher. Moreover, as the maximum temperature value at each time is mainly influenced by the power dissipated at that time, the effect of the first correction profile is predominant. These graphs, therefore, do not represent a clear proof that the scaling approach is a valid one. For this reason, also the temperature in a fixed location is reported as a function of time (Figure 5.22 (c)). These data, which are represented with the same legend as for Figure 5.22 (a), refer to a point whose distances from the edges of the die are 2 mm and 3.5

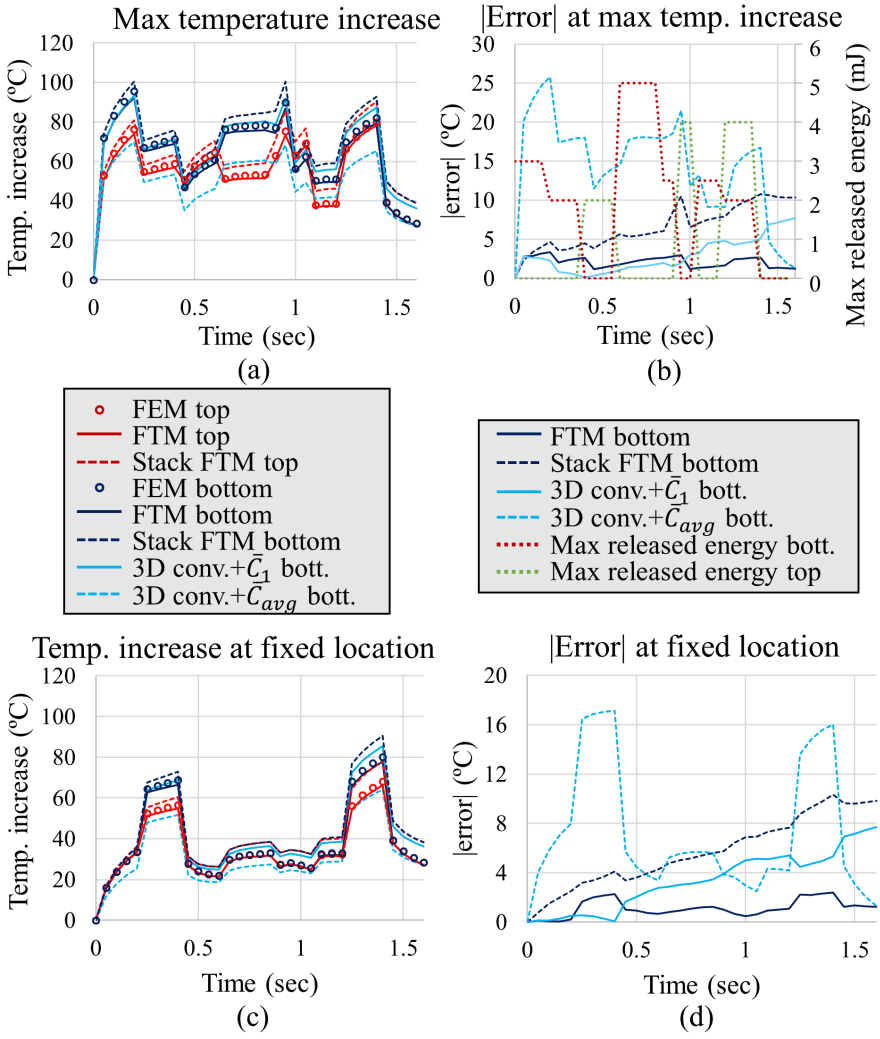


Figure 5.22: Results obtained, as a function of time, for a low power package configuration and time varying power maps. Different curves refer to results obtained with different methodologies and/or to different dies according to the legend. (a) Maximum temperature increase. (b) Absolute error in the location of the maximum temperature. On the right vertical axis, the maximum released energy is reported. The orange and green dotted curves refer to this axis. (c) Maximum temperature increase in a fixed location. (d) Absolute error in the same fixed location as in Figure (c).

mm. The graph demonstrates that the package correction methodology provides a significant improvement in accuracy everywhere, not only for the maximum temperature. In this way, the positive effect of subsequent correction profiles is proved.

Figure 5.22 (d) is the analogous of Figure 5.22 (b) but it refers to the fixed point results shown in Figure 5.22 (c). Similar comments as for Figure 5.22 (b) are appropriate here. The error referring to the stack FTM is, once more, increasing with time while, in the corrected model, this trend disappears. Since in this fixed location the temperature is mostly much lower than in Figure 5.22 (a), even if the error in absolute values is comparable to the one computed in the locations of the maximum temperature, it is much higher in percentage (after the first cooling down phase at 0.4 sec it is already around 20%).

Alternative package correction approaches

The error metric proposed in Section 5.4.3 gives an indication of the improvement, with respect to the transient stack FTM, achievable when the package thermal impact is included via the algorithm illustrated in this Section. However, there may be other options in between the stack FTM algorithm for the stack configuration and the corrected one for the package configuration. In some cases, a lower computational time may be more interesting than the higher accuracy achieved by the package FTM. For this reason, possible approaches, which place themselves in between the stack and the package corrected FTM both from an accuracy and from a computational time point of view, will be presented in this Paragraph.

The package corrected FTM, already presented in this Section, computes the temperature response of the system at a specific time t_k as the sum of impulsive temperature responses, $\bar{\Theta}_{z_i}(\cdot, \cdot, z_j, t_k; t_j)$. The whole power dissipation history is, indeed, assumed to be composed by subsequent impulses. The resulting impulsive temperature responses are individually corrected according to the time, with respect to t_k, the originating power has been dissipated. For this reason, the correction profiles are time dependent and a 2D-convolution approach is needed.

However, other methodologies can be considered in which the correction is applied a posteriori, on the final temperature profile at time t_k, without subdividing the power dissipation into its constituent impulsive components. In this way, 3D-convolution can be applied, with the advantage of shorter computational time. Since for the original package methodology, multiple, time dependent, correction profiles are used while, in this case, just one is needed, a choice on which one to consider has to be made. In the following, two options are proposed: an average, $\bar{C}_{avg}$, of the different $\bar{C}_k$, and $\bar{C}_1$. The former one doesn't give preference to any of the correction profiles. The latter one, on the other hand, considers just the correction referring to the end of the impulsive power dissipation. This is for the same reason explained in Section 5.4.3 (cf. Paragraph *"Temperature profile extraction"*) which is

that the latest dissipated power impulse normally has the highest impact on the final temperature profile. These simplifications are implemented to speed up the algorithm and, as a consequence, they neglect some of the physics underlying the phenomenon. What is ignored in this case is the fact that the thermal spreading is time dependent.

Results obtained with these algorithms are shown, for the bottom die, in Figure 5.22 with light blue curves. Full lines represent results obtained by applying $\bar{C}_1$ to the 3D-convolution final results, while dashed lines the results obtained by applying $\bar{C}_{avg}$. The figures prove that the approach based on the application of an average correction after 3D-convolution does not generate any improvement in accuracy with respect to the stack FTM. This is because, even if the shape of the correction profiles is almost always the same (cf. Paragraph *"Temperature profile extraction"* in Section 5.4.3), these profiles are scaled between different extreme values and, thus, the final applied corrections significantly differ from time to time.

The results referring to the application of $\bar{C}_1$ after 3D-convolution in Figure 5.22 (a) and (b), instead, show good accuracy. This mainly happens when the chip warms up; during the cooling down phase, after 1.4 sec, the accuracy is reduced. The other correction factors, not $\bar{C}_1$, should, indeed, be used in this situation since no power is being dissipated in this test case after 1.4 sec. It has to be noted that these curves refer to the maximum temperature, whose location varies during chip activity and follows the power dissipation location.

If the location is fixed, as in Figure 5.22 (c) and (d), the accuracy of the $\bar{C}_1$ correction approach deteriorates. This is, once more, due to the fact that, contrary to the previous situation, a fixed location experiences also cooling-down phases during chip activity. In this case the application of $\bar{C}_1$ after 3D-convolution is better than no correction at all but it is much worse than the full 2D-convolution method.

This means that, if the interest is just in accurately predict peak temperature during chip activity, then the application of $\bar{C}_1$ after 3D-convolution provides accurate results in shortest computational time. On the other hand, if more importance is given to the whole temperature profile, for which this high resolved FTM has been developed, the computationally more expensive 2D-convolution plus the time dependent correction methodology performs much better. However, in case of time constraints or reduced package impact, the algorithm based on 3D-convolution followed by $\bar{C}_1$ correction represents a good alternative.

5.4.6 Alternative computational approach: temperature only in selected points

Another possibility to reduce computational time is to limit the number of points in which the temperature is computed. If, for example, the interest is to remain below a threshold temperature and the locations where the high temperature increases

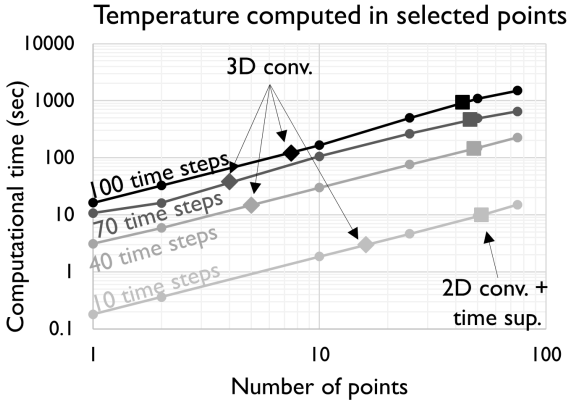


Figure 5.23: Dependency of the computational time on the number of points in which the temperature is computed for different simulated times. The computational time needed to obtain the full temperature map by applying the 2D-convolution+time superposition and the 3D-convolution approaches are also reported (squares and diamonds).

occur are known in advance, it is enough to follow the temperature evolution of these points. A slight modification of the developed FTM allows to calculate the transient temperature increase, with and without package correction, just in a selected number of points *without* affecting accuracy. All the information stored in the HSRs and in the PMs, which are used to compute the temperature increase in these points by the fully-resolved FTM, are, indeed, still used in the modified algorithm. The *point-FTM* works, indeed, by applying 2D-convolution (without FFT) just in the selected points. As a consequence, FFT cannot be implemented but the time dependent package thermal impact can be included.

The computational time related to the point-FTM can be much lower than in case of 2D- or 3D-convolution, depending on the number of points in which the temperature is computed, on the number of time steps and on the spatial resolution. Figure 5.23 presents a log-log plot showing the dependency of the computational time on the number of points in which the temperature is calculated for different simulated times. The data refer to a structure, discretized with a grid of 100×100 elements, with one power dissipation level and one temperature computation level. Three images per sides are considered. If the temperature is computed in just one point, the point-FTM is approximately two orders of magnitude faster than the 2D-convolution plus time superposition approach. In both cases the time dependent package thermal impact can be included. If the computational time of the point-FTM is compared with the one of the 3D-convolution, then the difference between having the temperature in a single point and having the full temperature map is reduced to one order of magnitude. In case of 3D-convolution, however, the time dependent package correction cannot be applied. The computational time

related to the point-FTM increases linearly with the number of considered points and it quickly becomes less convenient than using the 2D-convolution approach (around 50 points out of 10000). This is because, if the full temperature map is computed, then the FFT can be applied and this is not possible for selected points.

The point-FTM has a great potential in case of interest in few selected points because it drastically decreases the computational time without affecting accuracy. In case of FEM models, instead, even if the interest is to obtain the temperature only in one location, the result has to be computed in all the nodes in the model.

5.5 Summary

In this Chapter the steady state and the transient FTM methodologies for packaged 3D stacked ICs have been presented. They can be considered as multi-scale strategies whose core is constituted by a convolution based algorithm that allows the computation of the temperature increase, due to a generic, constant or time varying, power map in a stack configuration (stack FTM). The package spreading and capacitive effect is included via correction profiles. This is needed, in particular if the external thermal resistance of the package is high, because the stack FTM overestimates the temperature, especially in the corners of the stack. In case of the steady state regime, the correction profile is computed as the ratio between the thermal responses of the package and of the stack configurations to uniform power dissipation. In case of the transient regime, the time dependency of these thermal responses to uniform, impulsive power dissipation is accounted for. Moreover, in order to apply these corrections in transient simulations, 2D-convolution with subsequent time superposition has been implemented, instead of the less computationally expensive 3D-convolution, to obtain the temperature profiles.

The validation with respect to FEM of the full packaged 3D-IC shows that a significant improvement in accuracy is achievable by implementing this correction strategy. An error metric is also provided to allow the user to decide a priori if, for a specific situation, the relative achievable improvement is worth the higher computational time. Since the implementation of the package correction methodology in transient regime involves a significant increase in computational time with respect to the algorithm for the stack configuration, other strategies, which aim to implement the package correction while keeping a 3D-convolution based algorithm, are also presented and compared with the stack FTM and the 2D-convolution based corrected algorithm. The comparison shows that this last option provides a significantly higher accuracy all over the die. A modified algorithm, the *point-FTM* has also been presented. It allows to compute the package corrected temperature profiles just in few selected points with the same accuracy as the fully resolved model but much faster.

Chapter 6

Temperature Dependent Material Properties

6.1 Introduction

Let us consider a simple thermal conduction problem in which a one-material body is present, the heat path is unidirectional, no heat spreading occurs and uniform power is dissipated on one surface. The temperature increase due to the thermal conduction within this simple body is inversely proportional to the thermal conductivity k of the material of the body. This relation is for sure valid if k is constant (cf. Section 1.3.1); if this is not the case and the thermal conductivity is temperature dependent, $k = k(T)$, the relation between temperature increase and thermal conductivity depends on the function $k(T)$. It is nevertheless possible to state that the inverse proportionality relation is *approximately* valid. Starting from the Fourier's law

$$\boldsymbol{q} = -k(T)\boldsymbol{\nabla} T, \tag{6.1}$$

which for unidirectional heat flow can be written as

$$q_x = -k(T)\frac{dT}{dx}, \tag{6.2}$$

and integrating it

$$q_x l = -\int_{T_1}^{T_2} k(T)dT \approx -k(\bar{T})\Delta T, \tag{6.3}$$

the temperature difference between the active surface and another parallel surface in the body can be approximated as

$$\Delta T \approx \frac{l q_x}{k(\bar{T})} \tag{6.4}$$

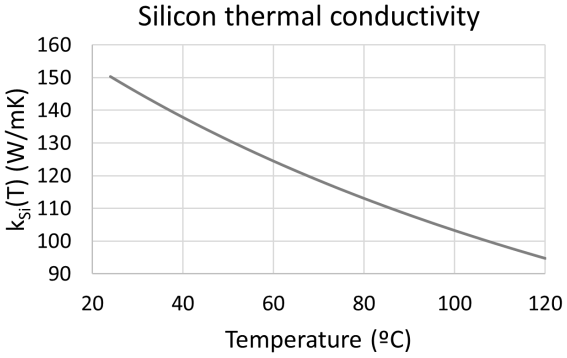


Figure 6.1: Silicon thermal conductivity as a function of temperature.

where q_x is the unidirectional dissipated power density, l the distance between the two considered surfaces, $k(T)$ the thermal conductivity of the material within them and $k(\bar{T}) = \frac{k(T_2)+k(T_1)}{2}$, with $\Delta T = T_2 - T_1$. As equations (6.3) and (6.4) show, the temperature difference between the two considered surfaces is, in particular, affected by the temperature dependency of the thermal conductivity. If, for example, k decreases with increasing temperature, the terminology *thermal runaway* is used: the temperature increase causes, indeed, a reduction of the k value, which causes an increase in temperature, and so on.

In microelectronics, this is of particular concern for silicon. The reason is twofold. First of all, higher temperature gradients are experienced in this material since power is dissipated *in* the silicon dies. Secondly, the dependency of the thermal conductivity of silicon, k_{Si}, on temperature is strong. More precisely, for the temperature ranges relevant for microelectronic applications, it is described by [82] (cf. Figure 6.1),

$$k_{Si}(T) = 148\left(\frac{300}{T + 273.15}\right)^{1.65} W/mK. \tag{6.5}$$

This means, for example, a reduction in thermal conductivity of almost 38% for a temperature variation from 28°C to 128°C. For the other materials in the 3D package this dependency is much lower (for copper, for example, it is around 1% for the same temperature variation [110]).

Since the partial differential equation governing the heat conduction phenomenon is

$$\rho(\boldsymbol{x})c(\boldsymbol{x})\frac{\partial T}{\partial t}(\boldsymbol{x},t) = \boldsymbol{\nabla}\cdot[k(\boldsymbol{x},T)\boldsymbol{\nabla}T(\boldsymbol{x},t)] + q(\boldsymbol{x},t), \tag{6.6}$$

the presence of materials with temperature dependent thermal conductivity makes this equation *non-linear*. This characteristic undermines the basic linearity assumption on which the FTM is built. To be able to apply superposition (and convolution), indeed, the governing equation has to be linear. In [90] the authors

studied the impact of non-linearity by comparing the temperature responses obtained by linear and non-linear RC-networks of a typical microelectronic package. They showed, in particular, that the non-linearity effect can be neglected if the difference between the maximum and the minimum temperature experienced during the considered phenomenon remains below $\sim 50°C$. If this holds, indeed, the error committed by using linear models is claimed to remain below 4%. Under this circumstance, therefore, the linearity assumption is acceptable and the convolution based FTM can be applied. However, in the same paper the authors also proved that, for higher temperature differences, the temperature dependency of the silicon thermal conductivity becomes significant and, as a consequence, it should be considered during modeling.

In [117] the authors propose two algorithms to include this non-linear phenomenon in their convolution based FTM. Both of them are iterative approaches based on look-up tables for the selection of the HSRs. The HSRs are, indeed, selected based on the k_{Si} value determined from the temperature computed in the previous step of the iterative algorithm. The difference between the two algorithms is that the first one considers a single HSR for each die, while, in the second approach, the HSRs selected at each iteration depend on the temperature of each specific grid point. For a $N \times N$ grid, N^2 different HSRs might, in principle, be needed. Both methods, however, introduce an increment both in computational time and in complexity because of their iterative nature and of the need of numerous HSRs.

In this Chapter, a *one-step* correction approach is presented to include the temperature dependency of the silicon thermal conductivity in the FTM. It is not based on iterations and just two HSRs have to be calculated for each specific amount of dissipated power. The basic steps needed in the algorithm to include both the package thermal impact and the temperature dependency of the silicon thermal conductivity are reported in the flowchart in Figure 6.2. The areas colored in gray, which represent the improvement with respect to the previous flowchart in Figure 5.2, highlight the small difference with respect to the algorithm that considered a constant value for k_{Si}. As it will be shown later on in this Chapter, however, the requirement to compute HSRs considering k_{Si} according to the dissipated power may significantly increase the complexity of the algorithm, especially in the transient regime. In this regime, in particular, a possible approach to tackle the non-linearity is reported in this Chapter, not a complete solution, together with the related remaining open issues.

6.2 Impact of $k(T)$ in the FTM

In order to highlight the error introduced in the FTM by the linearity assumption, a FEM study has been performed on simplified models. 2D-axisymmetric, parametric models have been considered for two dies stacks, in F2F bonding configuration, with power dissipated just on top of the bottom die. The modeling is performed

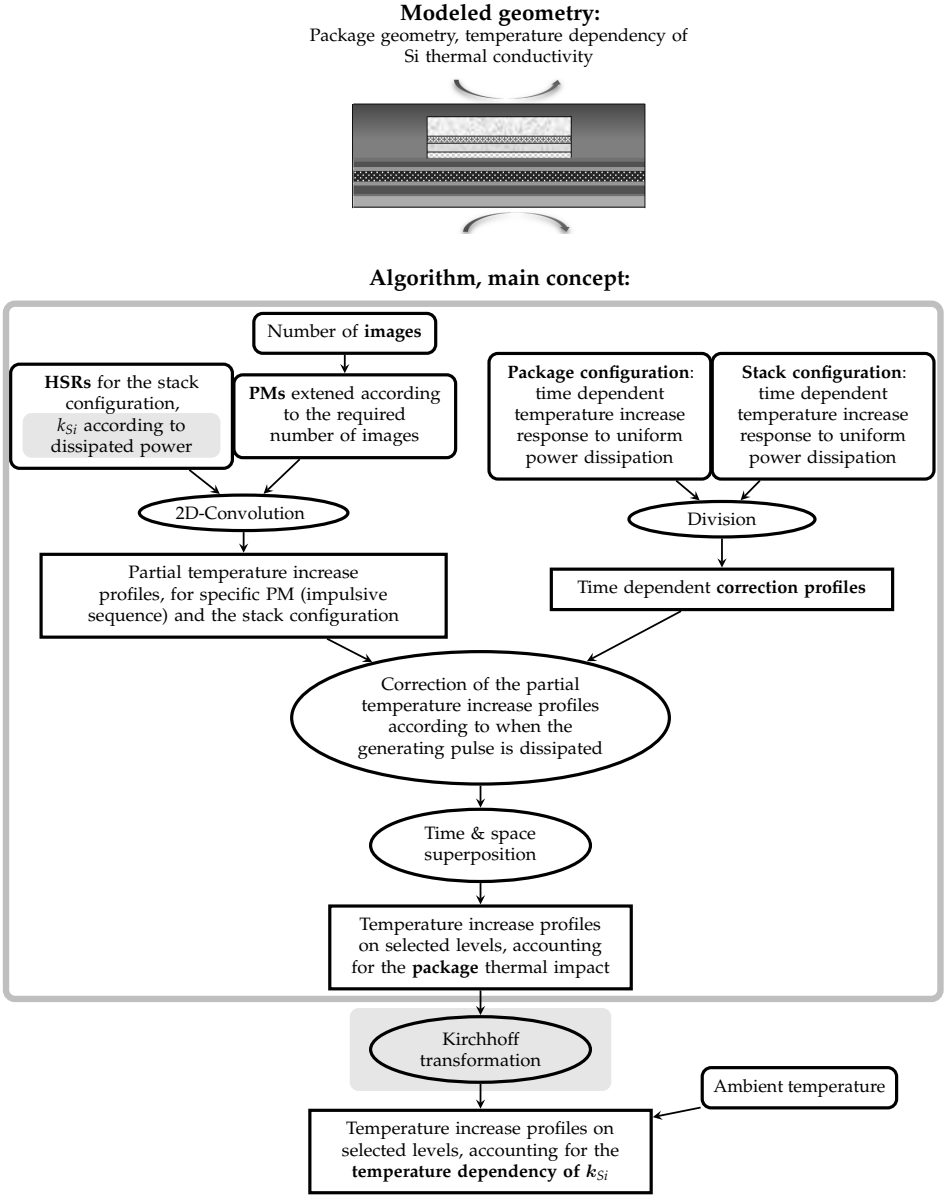


Figure 6.2: Modeled geometry and main concept of the algorithm described in this Chapter. The sections specifically introduced and discussed in this Chapter are highlighted in gray.

Table 6.1: Values use in the DOE to establish the impact of the temperature dependency of silicon thermal conductivity.

l_t (μm)	l_b (μm)	l_{interf} (μm)	k_{interf} (W/mmK)	h_t (W/mm^2K)	h_b (W/mm^2K)	radius chip (mm)	HS_{radius} (μm)	Q (W)
50	50	5	0.0002	0.0005	0.0005	4.08	120	0.4
500	500	20	0.005	0.02	0.02	8.16	360	0.8
						16.32	600	1.2
							1200	

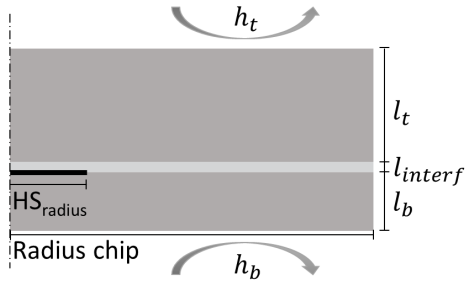


Figure 6.3: FEM setup for the DOE used to assess the importance of including the temperature dependency of k_{Si} in the FTM. The values of the considered parameters are reported in Table 6.1.

at the die stack level (the package is not considered) and hot spots of different dimensions are dissipated in the center of the bottom die. All possible combinations of the values of the design parameters reported in Table 6.1 have been considered, resulting in 2304 simulations. The FEM setup considered in the DOE is sketched in Figure 6.3. In each single case the maximum temperature obtained considering the temperature dependency of silicon thermal conductivity, according to equation (6.5), has been compared with the maximum temperature obtained assuming a constant value of k_{Si}. This last option represents what can be managed by the FTM.

The stack FTM works by convolving HSRs and PMs and, in order to compute the HSRs, a certain value of the silicon thermal conductivity has to be selected. Up to now this value has been considered to be independent of everything. Taking into account estimated working conditions, $k_{Si}(T)$ can be selected to be within $103 W/mK$, corresponding to 100°C, and $149 W/mK$ corresponding to ambient temperature. In Figure 6.4 (a) the error on the maximum temperature assuming a fixed value of $k_{Si}(T) = k_{Si}(68°C) = k_{Si} = 120\ W/mK$ is shown as a function of the temperature difference ($\Delta T = \max(T) - \min(T)$) in the active layer. The percentage error is computed as

$$\%err = \frac{\max(T_{k_{Si}(T)}) - \max(T_{k_{Si}=120})}{\max(T_{k_{Si}(T)}) - T_{amb}} \tag{6.7}$$

where $T_{amb} = 25°C$ while $T_{k_{Si}(T)}$ and $T_{k_{Si}=120}$ represent, respectively, the temperature values obtained considering temperature dependent silicon thermal conductivity

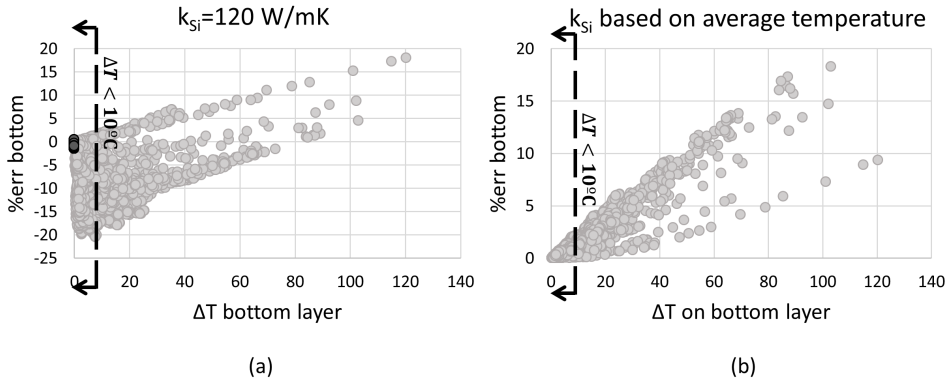


Figure 6.4: Percentage error on the maximum temperature introduced in the FTM if the temperature dependency of the silicon thermal conductivity is not taken into account for the 2304 simulations in the DOE of Table 6.1. The error is shown as a function of ΔT on the active die. In (a), a fixed value of $k_{Si} = 120\ W/mK$ is assumed while, in (b), the fixed k_{Si} value used in each simulation depends on the total dissipated power. The vertical lines indicate the areas for which $\Delta T < 10°C$.

and considering a fixed value of this parameter, $k_{Si} = 120\ W/mK$. Overall in this Chapter, the results are reported as temperature values, T, and not as temperature increases, Θ, because, since $k = k(T)$, the ambient temperature plays a role in determining $k(T)$. The gray circles in Figure 6.4 (a) are used for the results corresponding to all possible combinations of the parameters in Table 6.1 while the black ones for the results corresponding to uniform power dissipation for all the possible considered structures. In case the maximum temperature computed by the non-linear model is less than 68°C, $k_{Si}(T) > 120\ W/mK$ and, therefore, $\max(T_{k_{Si}(T)}) < \max(T_{k_{Si}=120})$: this explains the sign of the error in the graph. Two situations are of particular concern if the linear model with a fixed value of k_{Si}, independent of everything, is used:

- The average temperature increase differs a lot from the one the fixed value of k_{Si} refers to, which, in this study, is 68°C. These are, for example, the situations with a low amount of dissipated power and a large dissipation area, which are responsible for the negative error values on the bottom left of the graph;
- The HS power dissipation causes extremely high temperature differences ($\Delta T = \max T - \min T$) on the active layers, which result in the positive error values on the top right area of the graph.

Even if the graph is restricted to more common situations in microelectronic applications, in which the temperature differences ΔT on the active layers are more

limited (let's say 10°C), the error can still be higher than 20% if the selected k value strongly differs from the one corresponding to the average temperature of the chip.

A possibility to improve the temperature estimations, maintaining the actual working frame of the FTM, is to select the constant value of k_{Si} in the HSRs depending on the estimated average temperature increase in the chip. This means that, for each value Q of the total dissipated power in each die, a new set of HSRs has to be computed. The specific value of $k_{Si,Q}$, depending on the dissipated power Q, is calculated via resistance networks. A stack configuration is considered and the same total amount of power Q, dissipated on die i according to the power map PM_i, is assumed to be *uniformly* distributed over the whole active region. The specific value of $k_{Si,Q}$ is then retrieved from the calculated ΔT.

Figure 6.4 (b) shows the error calculated adopting this new procedure. As is clearly visible, a strong improvement with respect to the procedure with fixed $k_{Si} = 120\ W/mK$ is introduced: the issue related to the wrong selection of k_{Si} in the previous methodology disappears. The remaining high errors refer to situations in which the temperature difference within the active layer is high. If a restriction to more realistic situations with a limited temperature difference ($\max \Delta T = 10$°C) is considered, the error on the maximum temperature becomes less than 3%. Moreover, the error is always positive. This is because the selected $k_{Si,Q}$ values are computed assuming uniform power dissipation, which results in a lower temperature increase in the location of the maximum temperature than if the same power is dissipated in a HS. For this reason, $k_{Si,Q} > k_{Si}(T)$ in the location of high temperature and the maximum temperature increase computed assuming a fixed value of the silicon thermal conductivity is lower than in case the temperature dependent $k_{Si}(T)$ is considered. It has to be noted that the results concerning uniform power dissipation are not reported in the plot because $k_{Si,Q}$ has been selected so that the error in these settings is zero.

To summarize this second approach:

- The positive aspect is that a significant reduction in the percentage relative error is obtained;
- The negative aspect is that the HSRs depend on the amount of power dissipated in the specific configuration: for two cases with significantly different values of Q, two different sets of HSRs need to be computed.

6.3 Kirchhoff transformation

A possible solution to the remaining issue concerning the high temperature difference within the die is the use of the *Kirchhoff transformation* [3]. This is a technique that allows to linearize the heat conduction equation in the steady state regime.

Let's start defining the *apparent temperature* $\hat{T}$, which is a function of the *real* temperature T, as

$$\hat{T}(T) = T_0 + \frac{1}{k_0}\int_{T_0}^{T} k(\tau)d\tau \tag{6.8}$$

where T_0 is a convenient reference temperature and $k_0 = k(T_0)$ [3]. The application of this transformation to the heat conduction equation (6.6) results in

$$\nabla^2\hat{T} - \frac{\rho(\boldsymbol{x})c(\boldsymbol{x})}{k(\hat{T},\boldsymbol{x})}\frac{\partial\hat{T}}{\partial t} = -\frac{q(\boldsymbol{x})}{k_0}. \tag{6.9}$$

This transformed equation is still non-linear in the transient regime but, if it is restricted to the steady state regime, it is linear in the new variable $\hat{T}$

$$k_0\nabla^2\hat{T} = -q(\boldsymbol{x}) \tag{6.10}$$

and it can, therefore, be solved via the developed FTM.

By rewriting the dependency of silicon thermal conductivity in equation (6.5) in a more general form as

$$k_{Si}(T) = k_{ref}\left(\frac{T_{ref}}{T + 273.15}\right)^n \; W/mK \tag{6.11}$$

with $k_{ref} = 148\ W/mK$, $T_{ref} = 300\ K$ and $n = 1.65$, once $\hat{T}$ has been computed, the value of the original temperature variable T can be retrieved as

$$T = \left[\frac{(\hat{T} - T_0)(1-n)k_0}{T_{ref}^n k_{ref}} + T_0^{1-n}\right]^{\frac{1}{1-n}}. \tag{6.12}$$

It is important to note that T_0 and T_{ref} are not necessarily the same: the former is conveniently related to the transformation, while the latter is linked to the material property.

The application of this procedure in the FTM consists in an initial computation of the apparent temperature profiles $\hat{T}$ by convolving the HSRs, obtained using the value $k_{Si} = k_0$ for the silicon thermal conductivity, and the PMs. In a second step, the *real* temperature profiles T are retrieved by applying the *point-by-point* transformation in equation (6.12). A graph analogous to the ones shown in Figure 6.4 is presented in Figure 6.5. As can be seen from the Figure, the application of the Kirchhoff transformation further reduces the percentage error, which becomes less than 3.2% in absolute value even for really high temperature differences within the die. If the more realistic situation with $\Delta T < 10°C$ is considered, the error is kept below 1.3%.

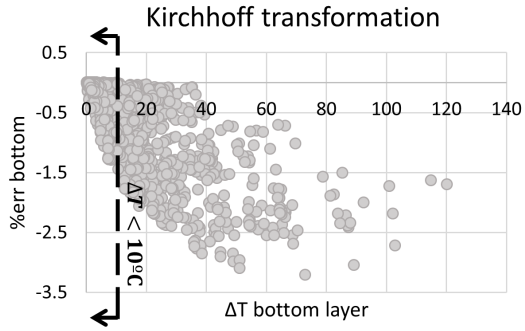


Figure 6.5: Residual error in the FTM after the application of the Kirchhoff transformation as a way to include the temperature dependency of the silicon thermal conductivity. The vertical line indicates the area for which $\Delta T < 10°C$.

6.3.1 Limitations

The residual error in Figure 6.5 is due to some constraints to the application of the Kirchhoff transformation that are not satisfied in the considered models and, more in general, in microelectronics applications. The corresponding limitations are listed hereafter.

Multiple materials: In the modeled stack configurations, multiple materials are normally present. The Kirchhoff transformation can be rigorously applied only if the thermal conductivity of *all* the materials varies with the same functional form, i.e. $k_i(T) = k_i f(T)$ with k_i dependent just on the material and $f(T)$ dependent just on temperature. If this is not the case, the continuity condition on the interfaces between different layers is not satisfied [10].

Convective BCs: Convective BCs are normally applied to the top and bottom boundaries of the modeled microelectronic device. The Kirchhoff transformation, however, has been mathematically derived for isothermal BCs and it cannot be rigorously applied in case of convective BCs. In [3] the authors proposed a partial solution to this issue, in case of *one* convective BC.

- For *uniform power* dissipation in a stack configuration, the corresponding uniform temperature increase on the boundary can be easily computed by means of corresponding R-networks. This means that the convective BCs can be transformed into an equivalent Dirichlet BC, for which the Kirchhoff transformation is mathematically valid. The value of T_0 in the transformation is selected as the temperature on the boundary.
- For *non-uniform power* dissipation, a proper selection of T_0 allows to use the Kirchhoff transformation in most situations typical for microelectronics, even if the transformation itself is not mathematically

valid anymore. This is possible because, in most situations of interest, the temperature on the boundary where convection occurs is quasi-uniform. The more the real situation deviates from this condition, the higher the error. Nevertheless, the value of T_0 should be selected in a specific way: it is the temperature of the boundary where convection occurs, as if the same power dissipated in the chip would be uniformly distributed on the active layers.

Moreover, in most cases of thermal modeling of microchips, convective BCs are applied on *two*, not on *one*, sides of the stack: top and bottom.

Package: This limitation is not related to the results in Figure 6.5, where only the die stacks are modeled, but it is related to the cases in which also the package is considered. The amount of material constituting the package can, indeed, be large and, if it is characterized by a high thermal resistance (plastic material, for instance), the temperature on the boundaries where convection occurs is much lower than the temperature *in* the silicon. This can represent a problem for the application of the Kirchhoff transformation if the *stack configuration* (and not the die stack) is considered in the stack FTM. This is due to three main reasons:

- The average temperatures on the two boundaries where convection occurs can be significantly different;
- The average temperature on the boundaries is significantly different from the average temperature in the silicon dies;
- The Kirchhoff transformation is applied to all the materials in the stack, also to the large amount of package material whose conductivity is not temperature dependent. This means that the thermal conductivity of all these materials is assumed to depend on temperature with the same functional form as silicon.

For these reasons, in case the difference between the k_0 values computed considering the average temperature on the surfaces of the stack configuration and on the surfaces of the die stack is higher than a user defined threshold ($3W/mK$, for instance), the temperature on the boundaries of the die stack is considered to compute k_0. These temperature values referring to the die stack are normally similar to each other because the die stack is mainly constituted by silicon, which is a good thermal conductor. This step is not against what was already explained for the application of the Kirchhoff transformation in case of non-uniform power dissipation and convective BCs. Indeed, if a configuration in which multiple, homogeneous material layers are stacked on top of each other with uniform power dissipation on one of these layers is considered, the thermal effect of the highest and/or lowest layers can also be included by removing these layers and adequately changing the heat transfer coefficients. This could be, therefore, a reason to use the die stack and not the stack configuration in the stack FTM. On the other hand, considering the full

stack configuration, instead of just the die stack, while computing the HSRs allows to have a better approximation of the package thermal spreading (cf. Section 7.4.2).

This means that, to obtain the results shown in Figure 6.5, the following assumptions have been made:

- The thermal conductivity of all the materials varies with the same functional form;
- The difference between the temperature on the top and on the bottom boundary of the die stack is not significant. As a consequence, it is assumed that the Kirchhoff transformation can be applied to problems with two convective BCs;
- Also related to the fundamental assumption of a limited temperature difference between the top and the bottom boundary, is the practical assumption that the T_0 value in the Kirchhoff transformation can be considered as the average between the temperatures computed on the top and on the bottom surfaces, as if the power would be uniformly distributed on the active layers. Normally, these top and bottom surfaces are the ones where the convective BCs are applied. In case of highly resistive packages, however, the top and bottom surfaces of the die stack are considered, even if more material is present around the stack and the BCs are not applied on those levels.

Since, as Figure 6.5 shows, the error reduces with respect to the previous approaches, these assumptions proved to be reasonable and the Kirchhoff transformation can be applied to improve the results of the FTM.

6.4 Steady state FTM including package spreading and $k(T)$

6.4.1 Flowchart of the FTM algorithm

In this Section, the various steps needed in the algorithm to include the temperature dependency of the silicon thermal conductivity in the FTM for a microelectronic package are listed together with the corresponding flowchart (cf. Figure 6.6). The updated steady state FTM includes, therefore, both the package thermal effect and the temperature dependency of k_{Si}. It consists of the following steps.

1. Computation of the package correction profile using $k_{Si} = k_{ref} = 148\ W/mK$ (cf. Chapter 5).

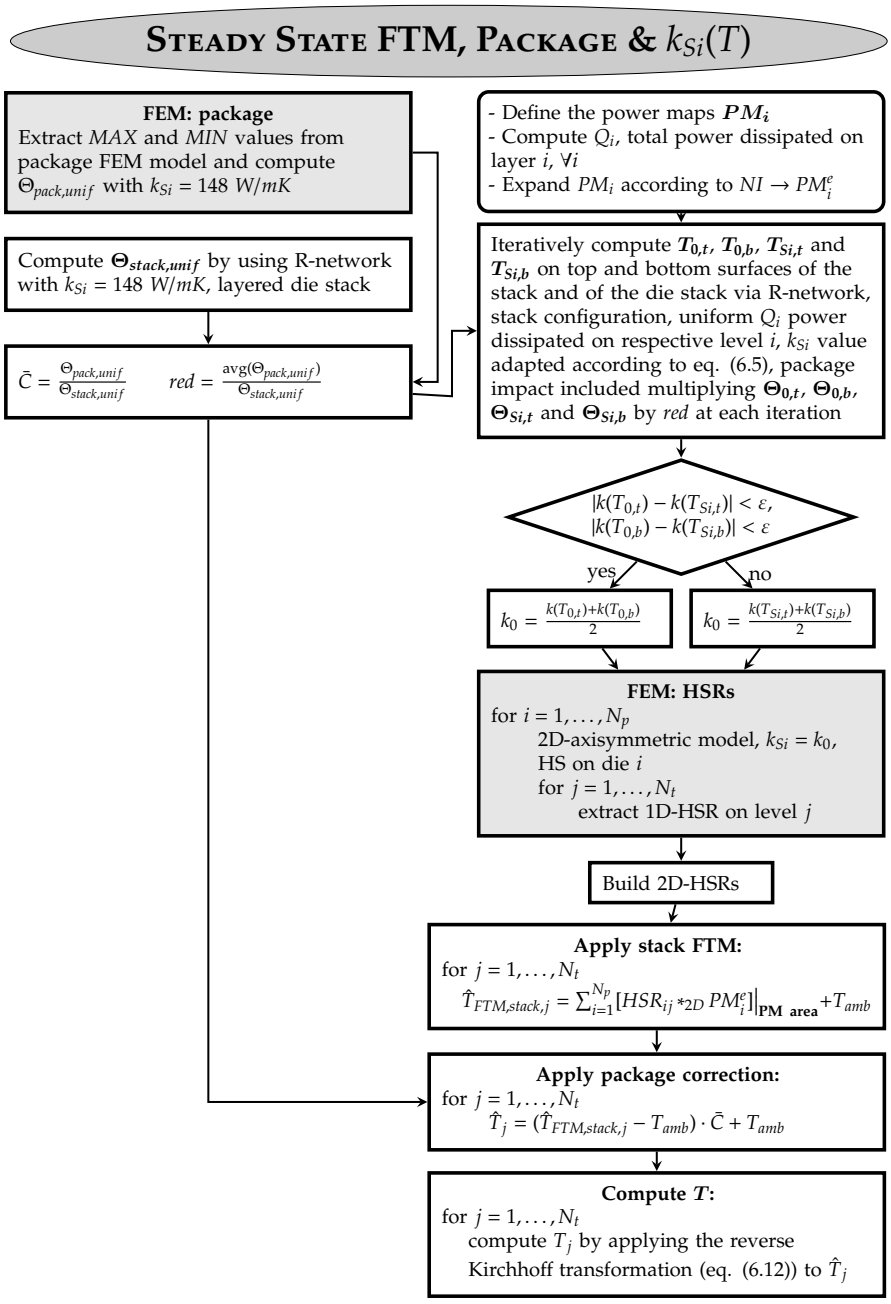


Figure 6.6: Flowchart representing the algorithm implemented for the steady state fast thermal modeling of packaged 3D-ICs, including the temperature dependency of the silicon thermal conductivity.

2. Calculation of the average reduction, due to the package impact, of temperature increase computed for a stack configuration according to

$$red = \frac{\text{avg}(\Theta_{pack,unif})}{\Theta_{stack,unif}}. \tag{6.13}$$

3. Computation of the average temperature on the top and bottom boundaries of the stack configuration, via appropriate resistance networks, assuming that the same amount of power, which is dissipated in the PMs, is uniformly distributed on each active layer. The value of k_{Si} is adapted iteratively, according to equation (6.5) where T is the latest computed value, taking into account both T_{amb} and the package impact. The average package impact, in particular, is included during the iterations by multiplying the obtained temperature increase values by *red*. At convergence, $T_{0,t}$ and $T_{0,b}$, the temperature values on the boundaries, are obtained. $T_{Si,t}$ and $T_{Si,b}$, the temperatures on the top and bottom of the die stack, need also to be computed.

4. Compute $k_{0,t} = k(T_{0,t})$, $k_{0,b} = k(T_{0,b})$, $k_{Si,t} = k(T_{Si,t})$ and $k_{Si,b} = k(T_{Si,b})$ according to equation (6.5).

5. If $|k_{0,t} - k_{Si,t}| < \varepsilon$ and $|k_{0,b} - k_{Si,b}| < \varepsilon$, where ε is a user defined quantity (in this thesis, $\varepsilon = 3W/mK$), then the impact of the package material is limited and $k_0 = \frac{k_{0,t}+k_{0,b}}{2}$. Otherwise, the temperature values computed for the die stack are used and $k_0 = \frac{k_{Si,t}+k_{Si,b}}{2}$.

6. Computation of the HSRs assuming k_0 as the fixed value for the thermal conductivity of silicon in the 2D-axisymmetric FEM models.

7. Computation of the apparent non-uniform temperature $\hat{T}$ by means of the stack FTM.

8. Application of the package correction on $\hat{T} - T_{amb}$ to include the package thermal spreading for the particular PM (cf. Chapter 5).

9. Application of the reverse Kirchhoff transformation in equation (6.12) to retrieve T.

6.4.2 Results

Figure 6.7 (a) reports the validation results of this algorithm in case of a LP package (cf. Figure 5.13 and Table 5.1 in Chapter 5) and a PM presenting a HS in the center. The HSRs are computed based on the die stack, the package material on top and bottom is not included. The orange curve is obtained by the package corrected FTM while the blue one by applying the algorithm illustrated for the Kirchhoff

transformation. In both cases, according to the theory explained in the previous Section, the selected value of k_0 is 132 W/mK. The red curve represents the FEM results, with $k_{Si} = k(T)$, against which the model is validated. As can be seen, the application of the Kirchhoff transformation significantly reduces the error on the peak temperature: from 5.9%, in case of the package FTM, to 1.4% in case of the Kirchhoff-package FTM. The last curve in the graph, the green one, represents the temperature profile obtained by applying the Kirchhoff transformation with $k_0 = 148$ W/mK, i.e. in case the value of k_0 is not selected according to the algorithm. In this case the error, even if both the package correction and the Kirchhoff transformation are applied, remains 5.1%. This last result highlights the importance of selecting an appropriate value for k_0.

Figure 6.7 (b) shows the temperature profile obtained for a more general PM (cf. Figure 7.6 in Section 7.4.1). The orange curve refers to the results obtained by selecting the appropriate k_0 value, according to the dissipated PM and following the algorithm presented at the beginning of this Section, to compute the HSRs. Just the package correction is, however, applied, not the Kirchhoff transformation. The blue curve is obtained by applying the Kirchhoff transformation on top of those results. As it is visible, the difference between the two curves is irrelevant. This is mainly due to the fact that the overall temperature difference in the computed temperature map is limited. The temperature range is, indeed, less than 7°C while in the previous example, where the Kirchhoff transformation had an impact, it was around 50°C. This means that the selected k_0 value, $k_0 = 120$ W/mK, is a good approximation of the real $k_{Si}(T)$ value everywhere, and the specific Kirchhoff transformation is not needed. If, however, the wrong value of k_0 is selected, the error can be larger (the green curve is obtained with $k_0 = 148$ W/mK).

From this analysis, it is possible to conclude that:

- The Kirchhoff transformation can be applied, even if some assumptions are undermined;
- The Kirchhoff transformation improves the accuracy in case of a high temperature difference within the die and highly non uniform power maps;
- In case of more uniform power maps, the results obtained using a properly selected value of k_0 can be comparable with the ones obtained by the Kirchhoff transformation;
- For a single specific case this algorithm is computationally as expensive as the previous one without the inclusion of the temperature dependency of the silicon thermal conductivity. The k_0 value to be used in the calculation of the HSRs can, indeed, be computed upfront, depending on the total amount of dissipated power, by a simple resistance network. Contrary to what has been presented in [117], no iterations are required. However, if multiple PMs have to be tested for the same geometry, since k_0 depends on the total dissipated power, it may be that multiple HSRs have to be computed.

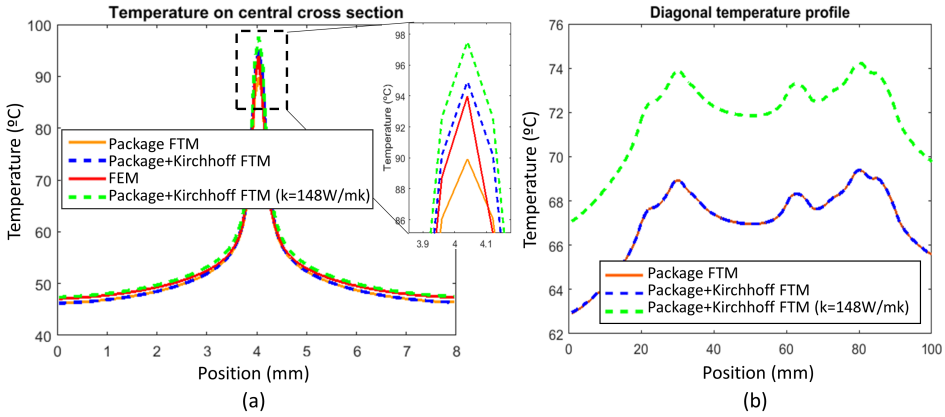


Figure 6.7: Validation of the algorithm to include the temperature dependency of k_{Si} in the FTM, steady state regime. Results refer to a LP package as described in Figure 5.13 and Table 5.1. The PM responsible for the results in Figure (a) has an HS in the center, while, in Figure (b), a more general, non uniform PM is used.

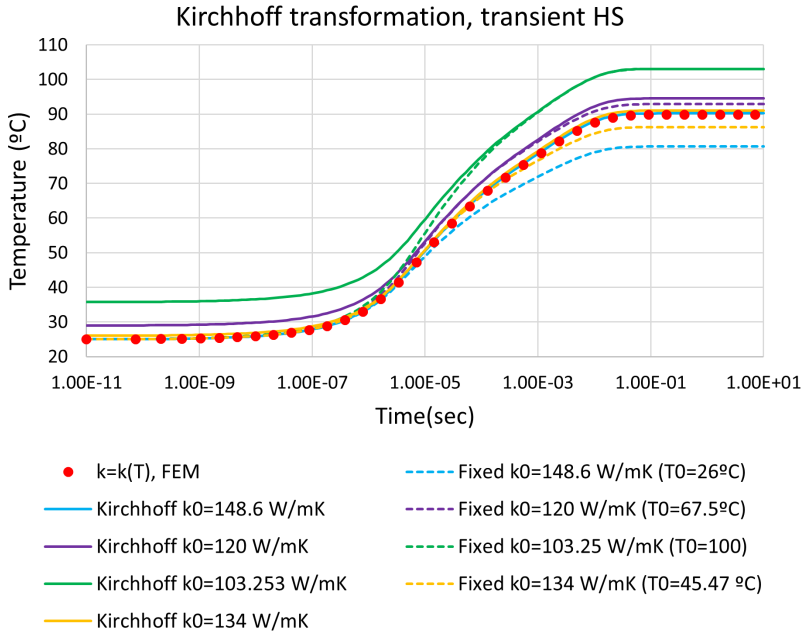


Figure 6.8: Kirchhoff transformation applied in transient regime for different values of k_0, HS power. The red dots refer to the FEM results obtained considering temperature dependent silicon thermal conductivity. The light-blue curves are obtained considering the k_0 value for the corresponding steady state regime. The other curves are obtained using other values of k_0. Dashed curves are obtained for specific k_0 values before Kirchhoff transformation, while full lines include Kirchhoff transformation.

6.5 Kirchhoff transformation in transient regime

In case of the transient regime, the transformed equation (6.9) is still non-linear. A further transformation has been proposed in literature to obtain the full linearization of the heat conduction equation in transient regime [5,6,47]. Defining a new *time* variable as

$$\tilde{\kappa}t' = \int_0^{t'} \kappa(\hat{T}(\tau))d\tau, \tag{6.14}$$

where κ is the thermal diffusivity ($\kappa = k/\rho c$), the time dependent heat diffusion equation becomes

$$\nabla^2\hat{T} - \frac{1}{\tilde{\kappa}}\frac{\partial\hat{T}}{\partial t'} = -\frac{q}{k_0}, \tag{6.15}$$

which is linear and can be solved by applying the FTM.

The issue at this point concerns how to calculate $\tilde{\kappa}t'$: it requires, indeed, the knowledge of how the apparent temperature changes over time. This means that, in order to know $\tilde{\kappa}$ and to calculate the HSRs with appropriate values of the material parameters, the temperature evolution of the system has to be known. This could be obtained, in principle, by running the FTM iteratively, meaning that the computational time drastically increases. However, in electro-thermal simulations, it is conventional to employ the Kirchhoff transformation and then to assume that $k(\hat{T}(t))$ in equation (6.9) is approximately constant [6].

To check if the approximation that employs just the Kirchhoff transformation is good enough, the algorithm has been tested for a couple of scenarios in which the impact of the non-linearity is large. Since the nonlinear term multiplies the time derivative, a case with a HS power dissipation (50 μm radius) and fast temperature increase has been tested. In particular a structure, consisting in just one silicon block with high convective BCs applied on both sides and without package, has been tested by means of a 2D-axisymmetric model. In Figure 6.8, the temperature evolution of the hottest spot of the chip is shown as a function of time. The red dots refer to the FEM results while the light-blue dashed curve to the ones obtained using the constant k_0 value properly determined by the total dissipated power. k_0 is computed following the algorithm presented for steady state and, therefore, the corresponding T_0 temperature is the steady state one. The profile obtained by applying the Kirchhoff transformation on top of this last result is represented by the full light-blue curve. In this case, therefore, even if just the Kirchhoff transformation is applied, without any transformation on the time variable, the error is already negligible.

The characteristic of this case is that the power dissipation is constant in time. In the transient regime, however, different amounts of power can be dissipated at different moments. This means that the T_0 value is time dependent: this is what the transformation in equation (6.14) takes into account. The other curves in Figure 6.8 show that, if the wrong k_0 value is selected, the error can be large.

The dashed curves refer to temperature profiles obtained for specific k_0 values but prior to Kirchhoff transformation, while the full lines are obtained after Kirchhoff transformation. The high error experienced in some of these cases means that it may be necessary to consider the *history* of the dissipated power while determining the k_0 value used to obtain the HSRs. This also means that it may be needed to use different HSRs at different time points in the transient FTM simulation, depending on the k_0 value referring to the temperature at that specific time point.

Moreover, a critical situation occurs if the average chip temperature drastically changes over time even for the same dissipated power and, therefore, for the same average steady state temperature. In the case shown in Figure 6.8, indeed, the power is dissipated in a small area, resulting in a strong HS (90°C peak temperature at steady state) but the average temperature remains almost the same from the beginning to the end of the simulation (25°C to 26°C). In this situation, therefore, if we would have considered the time dependency of k_0, $k_0 = k_0(T(t))$, then k_0 would have been almost constant.

In order to check a situation in which $k_0(T(t))$ strongly varies over time, a case with uniform power dissipation has been considered. The results are reported in Figure 6.9 for a stack configuration (cf. stack configuration for HP in Figure 5.4). In this case, the temperature in the chip at steady state is 129°C, corresponding to $k_0 = 91.4W/mK$. At the beginning of the simulation, however, the temperature is just 26°C, corresponding to $k_0 = 148.6W/mK$. The results obtained by considering $k_0 = 91.4W/mK$ and $k_0 = 148.6W/mK$, without applying the Kirchhoff transformation, are represented, respectively, by the blue circles and green crosses. Both of them are very close to the FEM result (red dotted curve), which includes the real temperature dependency of k_{Si}. This is because, in case of uniform power dissipation, the convective thermal resistance is much higher than the internal thermal resistance due to conduction (cf. Section 4.3.2). As a consequence, the selected value of k_0 has only a small impact on the final temperature. In case of HS power dissipation, however, the situation is different because the spreading resistance within the silicon stack plays an important role in the thermal phenomenon. As a consequence, for HS power dissipation, the proper assignment of the k_0 value is important to obtain accurate results. Going back to the uniform power dissipation scenario in Figure 6.9, if on top of the results obtained with fixed k_0 values the Kirchhoff transformation is applied, the full line curves are obtained. It is clear from the graph that the error strongly increases in both cases. If the k_0 value corresponding to the ambient temperature is selected (green curve), the obtained curve approximates really good the FEM results at the beginning of the simulation but the error is large when the temperature is high. The opposite occurs if the k_0 value corresponding to the steady state temperature at the end of the heating process is considered (blue curve). This means that the selection of k_0 should be time dependent, which would drastically increase the complexity and the computational time. However, if certain conditions are fulfilled and the applicability of the model is restricted to particular situations, it is still possible to sufficiently include the temperature dependency of the silicon thermal conductivity without too much computational

overhead. In particular,

- For *uniform power dissipation*, the impact of the value of k_0 assigned to the silicon material in the computation of the HSRs is not significant. For this reason, the dependency of k_{Si} on temperature can be ignored.
- For *hot spot power dissipation*, the application of the Kirchhoff transformation and the selection of a proper value of k_0 may strongly improve the accuracy of the results. To select k_0:
  - If $k_0 = k(T_{amb})$ and the Kirchhoff transformation is applied, a good approximation is obtained at the beginning of the simulation but, the larger the temperature increase during the simulation, the larger the error at steady state.
  - If $k_0 = k(\text{avg}(T_{ss}))$, where T_{ss} is the temperature at steady state, and the Kirchhoff transformation is applied, a good approximation is obtained at the end of the simulation but, the larger the temperature increase during the simulation, the larger the error at the beginning of the process.

  For really localized HS, $\text{avg}(T_{ss}) \approx T_{amb}$ and a good approximation is obtained everywhere.
- For cases *in between HS and uniform power dissipation*, the impact of the Kirchhoff transformation and of the selection of the value for k_0 depends on the ratio between the time step Δt used in the FTM and the time constant τ of the system. In particular, if the $\Delta t \approx \tau$ or $\Delta t \geq \tau$, the initial phase of the heating process in not included in the FTM. As a consequence, k_0 can be selected as $k(\text{avg}(T_{ss}))$ and the error is limited. In other cases, different strategies need to be applied but they won't be discussed in this thesis.

6.6 Transient FTM including package spreading and $k(T)$

6.6.1 Time dependent power maps

Let's consider a case in which the conditions for the application of the Kirchhoff transformation are fulfilled ($\Delta t \geq \tau$ or HS power dissipation) but the power map varies with time. As mentioned in the previous Section, in this case, it might be important to compute the HSRs taking into account the history of the dissipated power. This means, however, that we need to know, for all $t > t_k$, the effect of a certain amount of power dissipated at time t_k. This knowledge is equivalent to

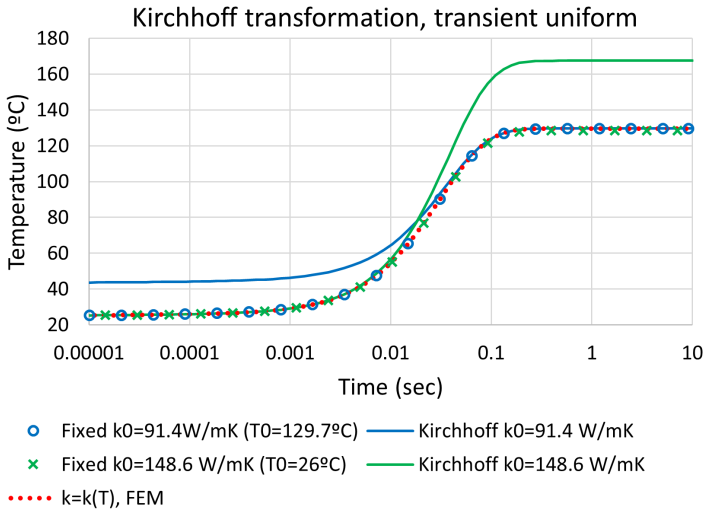


Figure 6.9: Kirchhoff transformation applied in transient regime for different values of k_0, uniform power. The red dotted curve refers to the FEM results obtained considering temperature dependent silicon thermal conductivity. Blue color indicates results obtained by considering $k_0 = k(T_{ss})$ while green color the ones obtained by considering $k_0 = k(T_{amb})$. Markers refer to the results before Kirchhoff transformation, while full lines include Kirchhoff transformation.

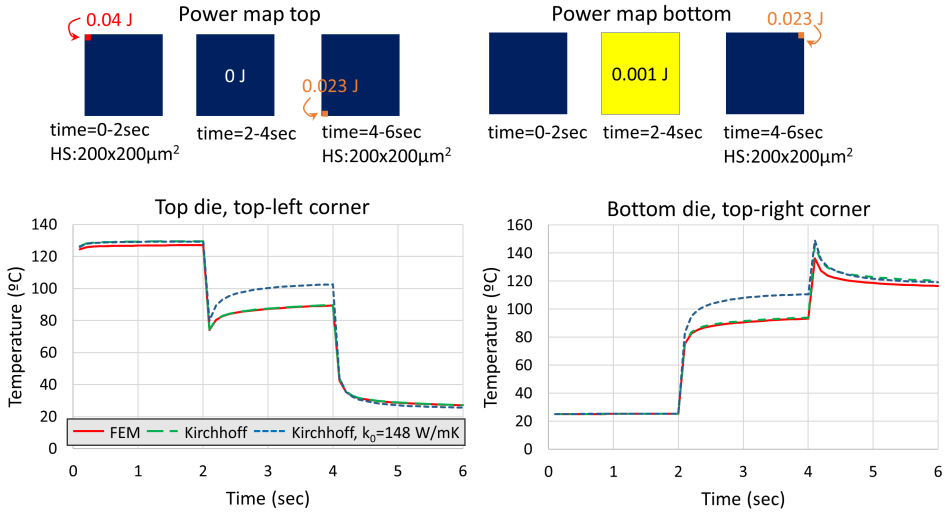


Figure 6.10: Kirchhoff transformation applied in transient regime for time varying power maps in case of a HP package as defined in Figure 5.13 and Table 5.1. The applied PMs and the temperature responses in the top-left corner of the top die and in the top-right corner of the bottom die are shown.

the one required for the application of the time transformation in equation (6.14). However, during chip activity, if the considered time step is larger than few μsec, the temperature in the die is mainly determined by the power dissipated at *that specific* time (similar reasoning as for the $\bar{C}_1$ package correction approach presented in Paragraph *"Alternative package correction approaches"* in Section 5.4.5). For this reason, just the *temporal sequence* of power maps is taken into account to compute appropriate k_0 values, without accounting for the history. This means that each single PM is considered individually, not as an element with a specific position in a temporal sequence. As a consequence, for each value of the total energy released during a single time step Δt, the $k_{0,Q(t)}$ value, corresponding to the steady state obtained for that specific power dissipation, is computed. Multiple HSRs are, then, calculated using these values. In order to avoid the calculation of a huge number of HSRs, the $k_{0,Q(t)}$ values are grouped in sets so that the difference, within the set, between the maximum and minimum $k_{0,Q(t)}$ value is less than 10%. The average of the $k_{0,Q(t)}$ values in each set is used to obtained the corresponding HSRs.

6.6.2 Results

Figure 6.10 shows the results obtained for a package configuration as the high power one presented in Figure 5.13 and Table 5.1 in Chapter 5. The algorithm used to compute the temperature profiles is an extension of the one reported for the steady state regime, just the computation of $k_{0,Q(t)}$ is different and it is performed as explained in the previous paragraph. These values are, in particular, time dependent and, as a consequence, the choice, based on $k_{0,Q(t)}$, of which HSRs to use in the convolution is also time dependent. The temporal sequence of the dissipated power maps on the top and the bottom die is shown on the top row of Figure 6.10 ($100 \times 100 \mu m^2$ spatial resolution). Since $\Delta t = 0.1 sec > \tau = 0.023 sec$, the Kirchhoff transformation with $k_0 = k(\text{avg}(T_{ss}))$ is implemented. The time evolution of the temperature on the top-left corner of the top die and on the top-right corner of the bottom die are presented in the left and the right graph, respectively. Red curves refer to FEM results, green curves are obtained considering the correct sequence of the $k_{0,Q(t)}$ values while, to obtain the blue curves, $k_0 = 148\ W/mK$ has been considered. In the last two cases the Kirchhoff transformation is also applied.

In this example there are three different sets of dissipated PMs. The $k_{0,Q(t)}$ value related to the first one, representing the energy released at each time step $t_k = \bar{t}_k \Delta t = \bar{t}_k \cdot 0.1 < 2$ sec, is 148.15 W/mK, i.e. $k_{0,Q(0)} = \ldots = k_{0,Q(1.9)} = 148.15\ W/mK$. For the other two sets of dissipated PMs, $k_{0,Q(2)} = \ldots = k_{0,Q(3.9)} = 107.21\ W/mK$ and $k_{0,Q(4)} = \ldots = k_{0,Q(5.9)} = 147.95\ W/mK$. However, since the difference between the first and the third set of $k_{0,Q(t)}$ values is less than 10%, the HSRs are computed just for two different k_0^l values, more precisely $k_0^1 = k_{0,Q(2)} = \ldots = k_{0,Q(3.9)} = 107.2\ W/mK$ and $k_0^2 = k_{0,Q(0)} = \ldots = k_{0,Q(1.9)} = k_{0,Q(4)} = \ldots = k_{0,Q(5.9)} = 148.05\ W/mK$. Moreover, since the selection of the HSRs is time dependent, a 2D-convolution based algorithm plus subsequent time superposition has to be applied. In particular, the selection of

the HSRs to be used in a specific 2D-convolution operation is performed according to the specific PM used in that operation. If the number of different HSRs is limited and the package correction is not included, however, the 3D-convolution algorithm can be applied multiple times and the obtained results are superposed afterwards. Each single run of the algorithm accounts for the PMs referring to a specific set of HSRs; the other PMs in the sequence are set to zero.

The obtained results clearly show that multiple HSRs are needed in case the released energy varies with time. The blue curves, obtained with a similar k_0 value as the one calculated for the first set of PMs, present, indeed, a high error in the center of the simulation, when uniform power is dissipated. The use of specific HSRs, which are calculated according to the specific released energy at each time step, shows higher accuracy.

This procedure improves, on the one hand, *accuracy* but, on the other hand, the complexity and the computational time needed to obtain the solution increase. The complexity, in particular, drastically increases for power maps that aren't periodic and in which the total amount of dissipated power strongly varies during chip activity. Moreover, as already mentioned for the steady state regime, the Kirchhoff transformation is relevant in case of a high temperature difference within the die. In the transient regime, the spatial temperature difference has to be considered together with the temporal variation in order to compute the HSRs considering appropriate $k_{0,Q(t)}$ values.

6.6.3 Flowchart of the FTM algorithm

Taking into account all the previous considerations, the flowchart in Figure 6.11 is obtained. Due to space limitation, the section corresponding to the package correction is not reported (cf. Figure 5.21). It is important to note that the decision block presents four options depending on if the package correction is included or not and on the number of different HSRs that have to be computed due to the variation of the dissipated power (i.e. the variation of k_{Si} for the average temperature increase at steady state). Moreover, as already mentioned before, the applicability of this approach is limited to cases in which $\Delta t \geq \tau$.

6.7 Summary

In this Chapter, a one-step algorithm to include the temperature dependency of the silicon thermal conductivity in the FTM has been introduced. It is based on the Kirchhoff transformation that allows, under particular conditions, to linearize the steady state heat conduction equation. In this regime, even if for the situations normally encountered in microelectronic packages the Kirchhoff transformation

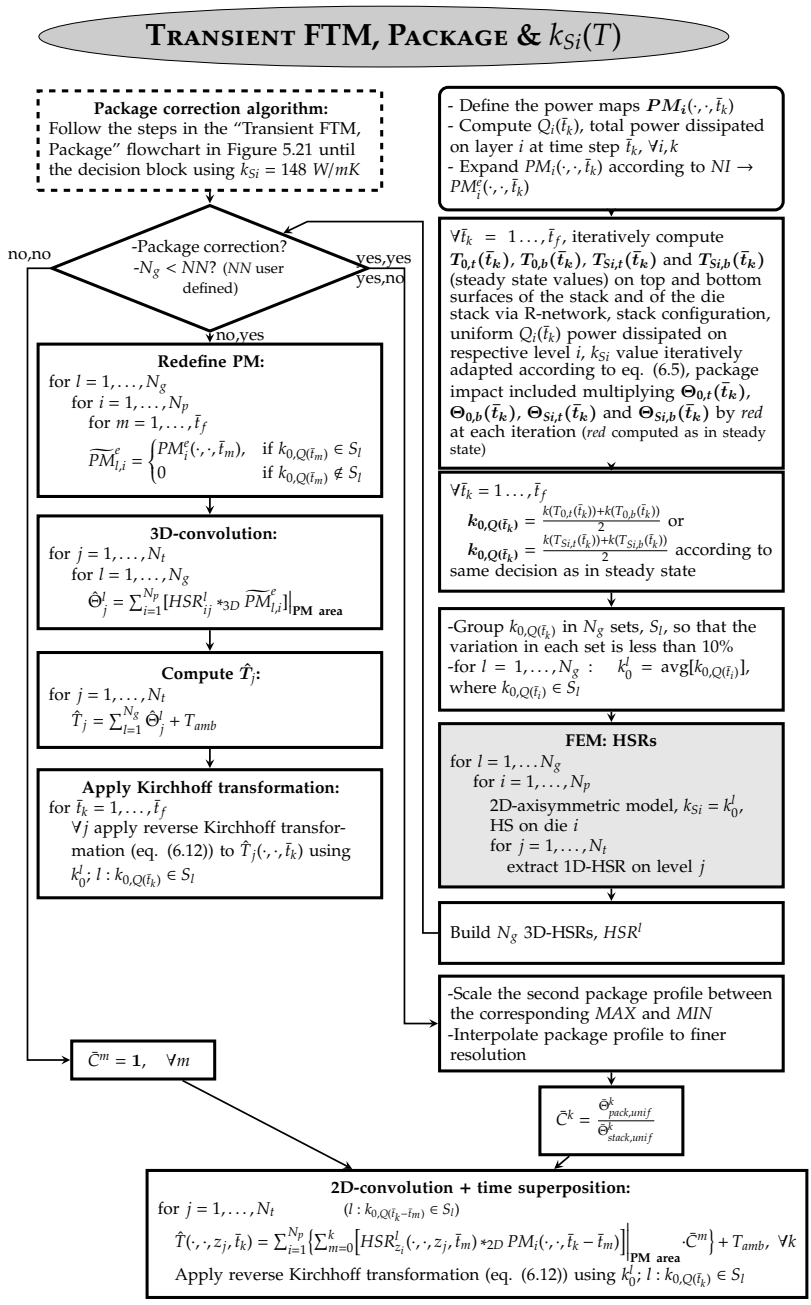


Figure 6.11: Flowchart representing the algorithm implemented for the transient fast thermal modeling of packaged 3D-ICs, including the temperature dependency of the silicon thermal conductivity.

doesn't rigorously linearize the heat conduction equation, the application of the Kirchhoff transformation improves the previous version of the FTM. To achieve this improvement, however, we have to pay on the re-applicability of the FTM. If the temperature dependency of the silicon thermal conductivity is taken into account, the HSRs need to be computed for appropriate, fixed values of $k_{Si,Q}$, which depend on the total dissipated power in the chip. The calculation of $k_{Si,Q}$ can be done upfront by means of a simple resistance network. Once this value is known, the HSRs are computed by means of 2D-axisymmetric FEM models without involving any iterative procedures. However, while previously, given a specific structure, the HSRs could be used to calculate the temperature increase due to *any* PM, this is not the case anymore: the HSRs are *power-dependent*. It is important to note that the selection of a proper value of $k_{Si,Q}$ depending on the dissipated power is always advisable (can be avoided for uniform power dissipation and packages with high external resistance). On the other hand, the application of the Kirchhoff transformation improves the temperature estimation just in case of a high temperature difference within the chip.

An algorithm to improve the results accounting for the temperature dependency of k_{Si} in the transient regime has also been proposed. It is valid only in specific situations, in particular if $\Delta t \geq \tau$, and further research has to be performed to make it more general. This algorithm is an extension of the one illustrated for the steady state regime. To define the value of k_0 used to compute the HSRs at a specific time step, the amount of energy released during that time step has to be taken into account. This means that, for a single simulation, multiple HSRs are needed and that the 2D-convolution plus time superposition based algorithm needs to be applied. This proposal doesn't solve completely the problem because it doesn't consider the *history* of the dissipated power. It improves, nevertheless, the accuracy in case of high temperature differences both in space and time and if $\Delta t \geq \tau$. In other circumstances this approach should be avoided. In case of uniform power dissipation, in particular, the proper selection of the k_0 value is not significant and the Kirchhoff transformation can be avoided.

Part III

Experimental Validation & Case Studies

Chapter 7

Experimental Validation

7.1 Introduction

"No one trusts a model except the man who wrote it; everyone trusts an observation, except the man who made it". This famous quote by Harlow Shapley nicely summarizes the scope of this Chapter. Up to now, the FTM has been compared with FEM results and it showed very good accuracy. This means that the proposed modeling methodology has been successfully validated with respect to another, well established, modeling methodology. Although FEM simulations are commonly accepted, they are already based on interpretation and simplification of the reality. As a consequence, they may suffer of lack of accuracy. For these reasons, in this Chapter, both the FEM and the FTM results are compared with experimental data for a packaged test chip. These data come from real devices and, as a consequence, they are not subjected to simplification and errors in the numerical approximation. They may, however, suffer from measurement errors and uncertainty.

In Section 7.2 the test chip is introduced and the setups used for the steady state and the transient validations are described in Section 7.3. In Section 7.4, then, the experimental validation of the FTM is presented for two package configurations in steady state and for three different power dissipation scenarios in the transient regime.

7.2 Test vehicle: PTCQ

A dedicated stackable test chip, named PTCQ (Packaging Test Chip version Q) and designed at imec, has been used to validate the model (cf. Figure 7.1). This is a $8 \times 8mm^2$ chip and it is equipped with specific structures that enable

electrical stress [15, 16] and temperature measurements [73–75]. These structures are included in basic cells of 240×240 μm^2, which are grouped in sixteen 8×8 arrays (cf. Figure 7.1 (a)), resulting, therefore, in a total of 32×32 basic cells in the chip. There are three types of cells (cf. Figure 7.1 (b)):

- CELL #1 has a global stress sensor and contains a diode as thermal sensor (pink color in Figure 7.1 (b));
- CELL #2 contains a heater and it includes a diode as thermal sensor (blue color in Figure 7.1 (b));
- CELL #3 has a local stress sensor as well as a diode used as thermal sensor (green color in Figure 7.1 (b)).

From a thermal point of view, therefore, this means that all the blue cells (75% of the chip area) can be used as heaters and that the temperature can be measured in all the 32×32 cells (type 1, 2 and 3), providing a temperature map with $240 \times 240 \mu m^2$ resolution. The heaters, in particular, can be switched on and off individually.

The diodes used for thermal measurements are located in the center of each cell and their calibrated sensitivity, in the range from 10°C to 75°C, is $-1.55 \pm 0.02 mV/°C$ for a current of $5\mu A$ (cf. Section 7.3 for more information). The heaters in cells #2 are constituted by two $200 \times 100 \mu m^2$ metal meanders resistors in the BEOL (cf. Figure 7.1 (c)) that are controlled independently by switches. As a consequence, each of these cells can have both meanders, just one meander or none of them active, resulting in a programmable power map ranging from a maximum of 75% coverage to localized hot spots of a single cell.

The PTCQ test chip has been designed is such a way that it can be used in 3D-stacking as well as in interposer configurations (cf. Section 8.2). In this Chapter, a PTCQ-on-PTCQ stack in a F2B configuration is considered. The top die is $200 \mu m$ thick while the bottom one is thinned down to $50 \mu m$ and it contains $5 \mu m$ diameter TSVs. The two dies are connected through Cu-Sn-Cu μbumps ($15 \mu m$ diameter on the top die side and $25 \mu m$ diameter on the bottom die side) with $40 \mu m$ pitch. This means that an array of 6×6 μbumps is located in each cell. The final thickness of the interface layer is $13 \mu m$ and underfill material, with $k = 0.4 W/mK$, is present between the μbumps. The layout of the μbumps is reported in Figure 7.2; the μbumps array covers $\approx$ 80% of the chip area (in correspondence of cells #1 there are no μbumps). The bottom die is connected to the substrate through $5 \mu m$ thick, $50 \mu m$ diameter, Cu pillars with a pitch of $170 \mu m$: in the area of each cell 2 Cu pillars are present to allow the current to reach the heaters in the two tiers (cf. Figure 7.2). The die stacks are packaged face down in a $14 \times 14 mm^2$ flip-chip ball grid array package (cf. schematic in Figure 7.3) with a $330 \mu m$ thick substrate. Different measurement environments and package configurations have been considered.

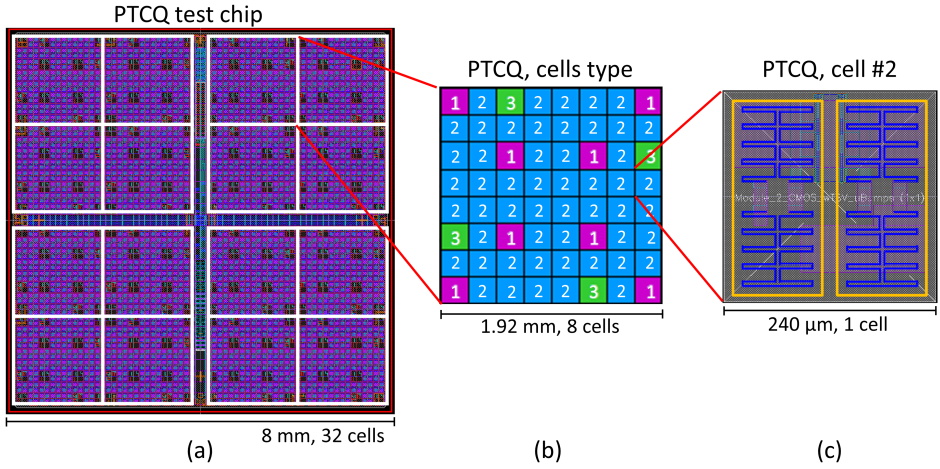


Figure 7.1: (a): Layout of the PTCQ test chip. (b) Organization of the different types of cells in basic modules. (c): Layout details of cell #2, the one with heater elements [73,75].

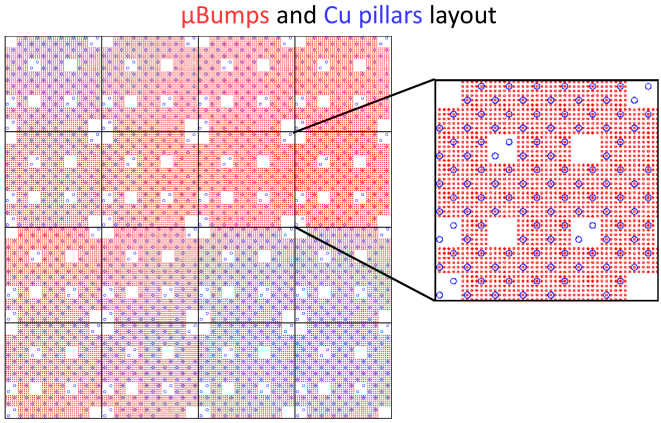


Figure 7.2: Layout of the μbumps (red) and the Cu pillars (blue) in the PTCQ-on-PTCQ stack (from [73]).

7.2.1 Low power package configuration

This first considered setup is used to mimic low power applications and it is illustrated in Figure 7.3 (a). An overmolded package is, in particular, considered. A schematic of this structure is shown in Figure 7.3 (a), together with a picture of the actual fabricated device. An epoxy mold compound, with a thermal conductivity value of $k = 1.0 \pm 0.1\ W/mK$ and a thickness of $0.7mm$, is used. The total thickness of

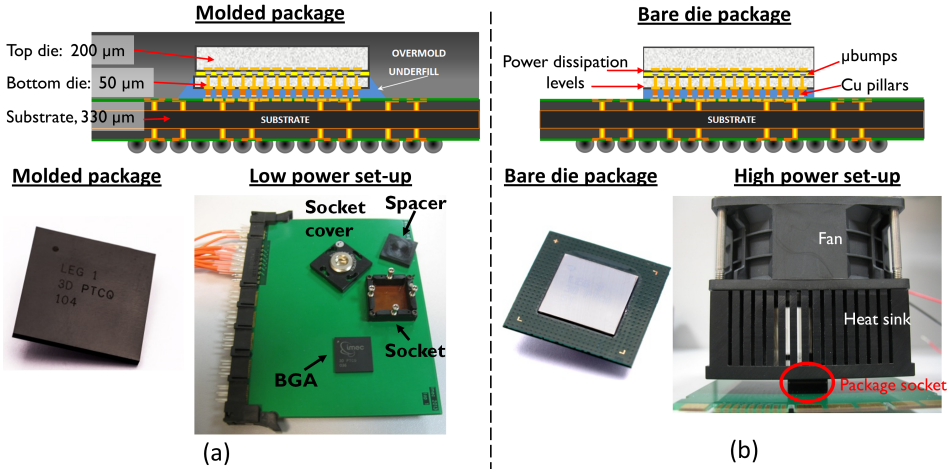


Figure 7.3: Schematic of the package, pictures of the fabricated packaged PTCQ and of the measurement setups for (a) the low power and (b) the high power configurations [73–75].

the package is 1.16mm. In order to perform measurements, this package is placed in a plastic socket that is attached to the PCB. A plastic spacer and the socket cover are, then, placed on top of the PTCQ package. The application of enough pressure, obtained by screwing the socket cover to the socket itself, allows a good electrical connection between the solder balls of the PTCQ package and the connections of the socket. There are three main features proper of this configuration:

- The PTCQ package is not permanently attached to the PCB. This allows to measure multiple test packages using the same measurement environment and the same socket;
- The materials used in the overmold and in the socket have low thermal conductivity. The whole setup has, therefore, high thermal resistance;
- Since no specific cooling solutions are applied, the main part of the heat is removed from the bottom of the package, through the PCB.

7.2.2 High power package configuration

To emulate a high power application, a different package and measurement configuration is considered (cf. Figure 7.3 (b)). The die stack is still placed in the measurement socket but it is not overmolded and a $80 \times 80 mm^2$ heat sink is directly attached to the backside of the top die. On top of it, a big fan is placed, providing forced convection cooling. In this way, the external thermal resistance is

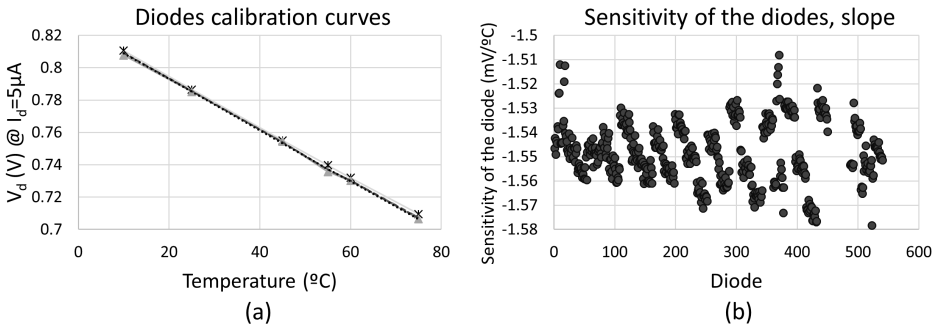


Figure 7.4: Calibration curves and sensitivity of the diodes in the PTCQ test chip for a current of $5\mu A$.

much smaller than in case of the low power package configuration and the heat is mainly removed from the top side of the stack.

7.3 Measurement procedure

Temperature measurements are performed by measuring the forward voltage drops in the diodes at $5\mu A$ current. Figure 7.4 shows the calibration graphs for the temperature sensitivity, σ, of the diodes in the PTCQ test chip. Graph (a), in particular, presents the calibration curves, for a current of $5\mu A$, for five different diodes in the PTCQ. As it is clearly visible, the linear relationship between temperature and voltage is approximately the same for all the represented diodes. This is true also for the diodes that, for readability reasons, are not shown in this graph. The results obtained for the calibration curves are, indeed, very stable and a very small variation, always within the measurement error, is detected. Figure 7.4 (b) shows the temperature sensitivity ($mV/°C$ @ $5\mu A$) for more than 500 different diodes. All the obtained σ values are very similar to each other and centered around an average of $\sigma = -1.55 mV/°C$ (95% confidence interval of ±1%). Since all the diodes respond *in the same linear way* to a change in temperature, the temperature increase ΔT, with respect to the initial temperature before power dissipation, in each diode can be computed based on the difference in voltage, ΔV, measured in each location with respect to an initial voltage, measured before power dissipation. Independently on where ΔV is measured, to obtain ΔT, it is enough to divide that value by the same sensitivity value σ. This is justified in the temperature range in which the calibration has been performed (cf. Figure 7.4 (a)) and for the same low current ($5\mu A$) in the diodes.

Concerning the power supply, as already mentioned, the two metal meanders in the heaters are controlled independently by switches. Due to the use of these switches, the voltage that can be supplied to each cell is limited to 1-1.5 V. Since each

heater has a resistance of 10Ω, this results in $100mW$ maximum power dissipation per cell, or $83.2W$ in case of all cells active. This means that, to activate all cells with this high power, the high current level of $83.2A$ is required.To estimate the losses in the PCB due to Joule heating and to know exactly how much power actually reaches the dies, 4 point measurements are performed during power dissipation.

All the measurements presented in this Chapter have been defined in discussion with dr. ir. Vladimir Cherman and kindly performed by him.

7.3.1 Steady state measurements setup

Steady state measurements are performed for both the LP and the HP setup. In both cases, data concerning different diodes are read out *sequentially* through an analogous shift register. To obtain the full temperature maps, all the diodes are scanned, column by column, starting from the bottom right corner. This operation, for all the 2048 sensors in the two dies, takes more or less 20-25 minutes and it allows to obtain two temperature maps with a resolution of $240 \times 240\mu m^2$ (as the ones shown in the first column of Figure 7.7, for example). Since these are steady state measurements, in order to make sure that the steady state condition is met, it is important to wait a sufficient amount of time between the beginning of the power dissipation and the beginning of the temperature readings. This is especially the case for the LP package (cf. Figure 7.5). In the considered cases, the waiting time is 30 minutes.

It is also important to note that, since to activate all the cells in a PTCQ test chip a current of $83.2A$ is required, a high current power source needs to be used. In particular, a current power source with a maximum of $100A$ is used in these steady state measurements.

7.3.2 Transient measurements setup

The transient measurements that will be shown in this Chapter refer to the low power package configuration. In general, transient analysis presents numerous advantages with respect to steady state.

- It allows to monitor the thermal delay between different dies and/or between different locations in the same die;
- Since the heat reaches the different levels and materials of the package with different delays with respect to the beginning of the power dissipation, performing a transient analysis allows to calibrate the models more accurately, by distributing the thermal resistances over the different materials, and to extract more accurate information on the material properties.

The PTCQ test chip was, originally, designed for steady state measurements. It is, nevertheless, possible to perform transient measurements with some limitations that come from the design of the test chip itself, from the measurement equipment and from the implemented test procedure. The two main issues faced for transient measurements are:

- High currents and fast switching times can create very high peaks of voltage across parasitic inductances. This is because $V = L\frac{di}{dt}$, where V is the voltage, i is the current and L the parasitic inductance. These voltage peaks cause difficulties to understand to which extend the measured voltage drops are due to temperature variations.
- The high current power source used for steady state measurements is very inert and can not be used to generate fast transient voltage steps. To analyze the short time transient responses and, therefore, the thermal properties of the materials close to the power dissipation location, a fast current switching is required.

For these reasons, a low voltage, high speed data acquisition card and a DC-AC voltage amplifier are used instead of the high current power supply for fast transient measurements (down to $10\mu s$). Moreover, there are additional voltage drops in the setup and, as a consequence, the power provided to the heaters can be even smaller than the one originally supplied. For this reason also the voltage in the heaters, not just the one in the diodes, is measured during the experiments. The maximum current that can be supplied to the PTCQ in the transient setup is $1A$ and, since the voltage per cell is limited to $1 - 1.5V$ and, as a consequence, $100mW$ are dissipated per cell, a maximum of approximately 10 heaters can be activated. More heaters can, in principle, be switched on but this will reduce the supplied voltage to the cells.

Moreover, only two different types of power maps can be considered in the implemented test procedure for a single measurement: one with all heaters off and another one in which a group of heaters is switched on. These two PMs can be alternated multiple times but a third PM, with a different group of active heaters, cannot be considered in the same measurement using the current test procedure.

Concerning the temperature measurements, in the steady state setup, data concerning different diodes are read out *sequentially* through an analogous shift register. Simultaneous reading of more sensors is not possible and, since the aim of the transient measurements is to record the time evolution of the device with high temporal resolution, just one sensor is monitored in each measurement. If the temperature has to be monitored in different locations for the same dissipated PM, multiple measurements have to be run.

Finally, in the implemented test procedure, the time resolution of the measured data depends on the duration of the whole experiment. In each measurement,

the temperature is recorded at 200000 equidistant time points. This means that information concerning the very short time and the very large time system response cannot be collected in one single measurement. Short measurements (2 sec) are run to obtain the information in the time interval close to the starting of the power dissipation (10 μsec). These results provide information concerning the materials close to the heat generation region (normally the die itself and the interface layer). Longer time measurements (up to 3600 sec) are, instead, run to collect information about the sections of the package further away from the active region (normally the package and the boundary conditions). The combination of these data provides the full heating up (or cooling down) curve of the device. This is illustrated in Figure 7.5, which also shows that a long time is required to reach the steady state regime in this LP package configuration.

To summarize, the practical limitations encountered for the setup and the test procedure of the transient measurements are:

- Just a small number of heaters can be activated at the same time;
- The temporal change of the PM is limited to the switching (on-off and off-on) of a fixed group of heaters (multiple time switching is possible);
- In each measurement, just one diode can be monitored;
- Data are collected at a fixed time step;
- 200000 data points are collected in each measurement.

Despite these limitations, it is still possible to define useful test cases to validate different aspects of the developed FTM.

7.4 Experimental validation of the FTM

7.4.1 Modeling information

In this Section, the experimental validation of the FTM, both in the steady state and in the transient regime is presented. The dimensions and the material properties of the different modeled parts of the PTCQ-on-PTCQ stack are reported in Table 7.1. It has to be noted that the mold compound *encapsulates* the die stack (its thickness is reported starting from the top of the top die but, in the area around the stack, the mold reaches the substrate material) and that it is included just in the package considered in the low power configuration. Moreover, for some layers (BEOL, interface layer, Cu pillars) equivalent material properties are used. A FEM analysis of the same situations has also been performed and the results are reported in the

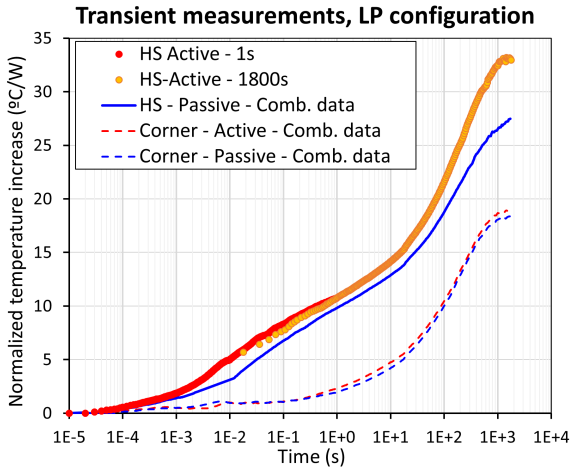


Figure 7.5: Processed measurement data reporting the full transient temperature evolution (combination of short and large time ranges) in different positions on the dies. HS power dissipated in the center of the bottom die and temperature measured in the center and in the corner of both the top and the bottom dies.

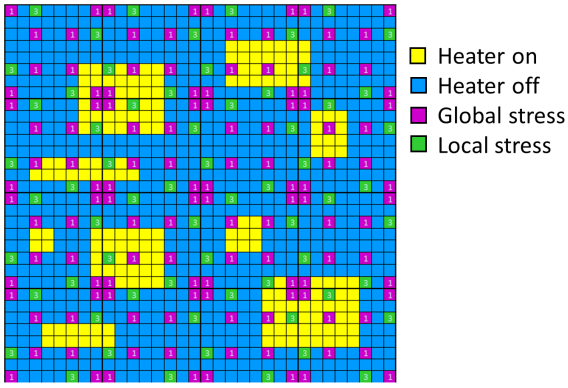


Figure 7.6: Dissipated power map on the bottom PTCQ die in the experimental validation of the FTM for the packaged 3D-stack in steady state [75].

following of this Section. More details about the FEM model itself are reported in Appendix A.1.

The PCB and the socket are included in the model by means of properly selected convective boundary conditions, they are not directly modeled. For the PCB, in particular, in order to include its capacitive impact in the transient regime, the substrate layer has been divided into two parts. Both of them have the same value of the thermal conductivity k and of the mass density ρ, but different values for the

Table 7.1: Values of the parameters used in the modeling of the PTCQ test chip. Dashed lines indicated the levels where the temperature is computed and point lines the levels in which power is dissipated. (*The thickness of the overmold is reported starting from the top of the top die.)

	Dimensions, $x \times y \times z$ (mm)	Thermal conductivity, k $(W/m\ K)$	Mass density, ρ (kg/m^3)	Specific heat capacity, c (J/kgK)
Overmold	$13.6 \times 13.6 \times 2.2^*$	$k = 1$	1120	1600
Top die	$8 \times 8 \times 0.2$	$k = k(T)$	2330	725
Top BEOL	$8 \times 8 \times 0.002$	$k_{x,y} = 0.2$, $k_z = 2$	2000	1464
Interface	$8 \times 8 \times 0.013$	$k_{x,y} = 0.5$, $k_z = 4.2$	3500	700
Bottom die	$8 \times 8 \times 0.05$	$k = k(T)$	2330	725
Bottom BEOL	$8 \times 8 \times 0.005$	$k_{x,y} = 0.2$, $k_z = 2$	2000	1464
Cu pillars	$8 \times 8 \times 0.1$	$k_{x,y} = 0.4$, $k_z = 2$	3500	770
Substrate1	$13.6 \times 13.6 \times 0.125$	$k_{x,y} = 12$, $k_z = 0.6$	1500	9600
Substrate2	$13.6 \times 13.6 \times 0.375$	$k_{x,y} = 12$, $k_z = 0.6$	1500	129528

specific heat capacity c. In this way, the steady state simulations are not affected by this division but, while dealing with the transient regime, the capacitive impact of the PCB is partially included without adding further complexity in the model. This is indicated by *Substrate1* and *Substrate2* in Table 7.1.

In Chapter 4 the stack FTM has been extended to include the thermal impact of specific μbump layouts. In this experimental validation, however, a homogeneous interface layer is considered and the equivalent material properties calculated for the μbump array (cf. Appendix A.1) are assigned to it. This is because $\tilde{\rho} = 0.8$ and, for such a large area of the interface layer covered by the μbump array, the approximation of considering an area array configuration is acceptable. Moreover, the value of $\tilde{\rho}$ is much larger that the maximum value considered in the fitting procedure performed in this thesis.

Another detail to be stressed is that the heaters are placed *in* the BEOL layers. For the steady state simulations, the decision of modeling the active layers on top or on bottom of the BEOL layers doesn't influence the results. However, for the very small time range in the transient simulations, it has been demonstrated that placing the active regions below the BEOL (flip chips, not in contact with the silicon) provides a better agreement with the experimental results. This is

Table 7.2: Values of the heat transfer coefficients and of k_0 used in the modeling of the PTCQ test chip, LP configuration.

	h_t (W/m^2K)	h_b (W/m^2K)	k_0 (W/mK)
FTM, HSRs stack conf., k_{Si} in stack	1631.15	672.68	120

because, the heating delay, from the power dissipation layer (below BEOL) to the temperature computation layer (bottom of the silicon), is accounted for in the HSRs.

7.4.2 Steady state regime

In both the low power and the high power configurations, the same group of cells is activated (cf. Figure 7.6). This is a non-uniform PM applied on the bottom, thin die of the stack. Multiple HSs of different sizes are considered in this PM, in order to obtain a trustworthy validation of the modeling methodology. In particular, 171 heaters (out of 832, the 20.5%) are activated on the bottom die while the top die is kept passive. Although the dissipated power map is the same in both the LP and the HP configurations, the amount of dissipated power is different. The forced cooling allows, indeed, to dissipate much more power for approximately the same maximum temperature increase. For the HP socket, in particular, 15.3W (1.5V, 10.2A) is dissipated while, for the LP socket, just 2.25W (0.5V, 4.5A) [75].

Low power configuration

In this Subsection, the model is validated with respect to the experimental results from the low power setup: molded package in the plastic socket with natural convection.

As already mentioned in Chapter 5, the package impact in this LP configuration is quite relevant (sp = 6.85% from equation (5.10)). As a consequence, in order to have a good estimation of the package thermal impact, the HSRs are computed starting from the *stack configuration*, including also the mold compound and the substrate on top and bottom of the die stack (cf. Figure 5.4 in Chapter 5). This means that a large amount of material is considered on top and bottom of the die stack itself. Moreover, the mold compound has a low thermal conductivity value. Due to these characteristics, there is a significant difference between the average temperatures computed, for the applied power map (Figure 7.6, 2.25W), on the top and bottom surfaces of the stack configuration at steady state: $T_{0,t}$ = 34.3°C and $T_{0,b}$ = 51.8°C (steps 1-4 in the algorithm presented in Section 6.4.1). According to step 5 in the same algorithm, the average temperature increases on top and bottom of the die stack itself are considered to compute k_0. Table 7.2 reports the values

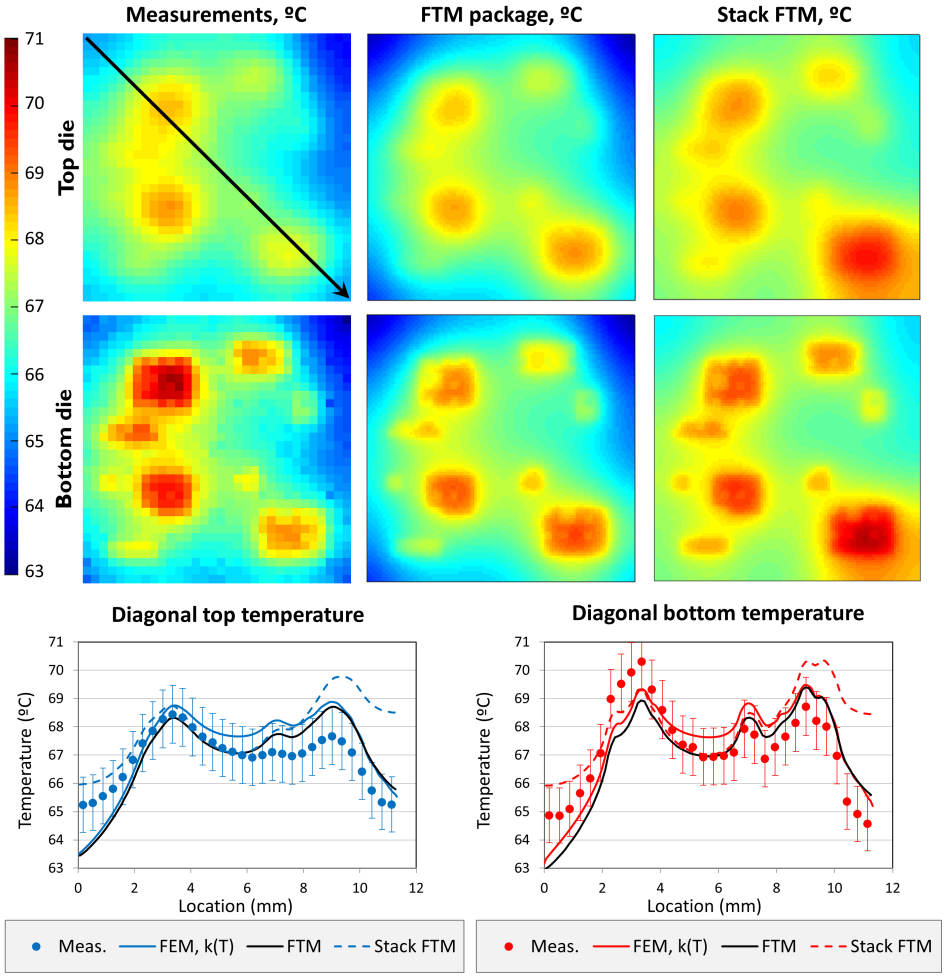


Figure 7.7: Temperature results obtained by measurements, by the FTM with package correction and by the stack FTM without package correction for the LP, PTCQ-on-PTCQ configuration. The diagonal cross sections allow for an easier comparison of the results.

of the applied heat transfer coefficient on top and bottom of the model for the HSRs and the k_0 value assigned to silicon. For a sketch of the considered geometry, please refer to the details concerning the low power configuration in Figure 5.13 while, for more information about the corresponding FEM model, to Appendix A.1.

The results are shown in Figure 7.7. The first and the second row refer, respectively, to the top and the bottom die. The first column shows the temperature maps

obtained by the measurements, the second one the data computed by the FTM including the package correction and the third one the results obtained by applying the stack FTM without package correction (T_{amb} = 25°C). The temperature cross sections, along the diagonal indicated by the arrow in the Figure, are reported in the third row for both the top and the bottom die. This allows an easier comparison of the results. In these two last graphs the measurement data are indicated by colored, circular markers, the stack FTM by colored, dashed lines, the complete FTM by black, full lines and the corresponding detailed FEM results by colored, full lines. In all these graphs, a strong thermal coupling between the top and the bottom die is visible and the pattern of the dissipated power map can be read, even if highly smeared out in the top die, on both chips.

From the results obtained by the stack FTM, the importance that has, in this scenario, the application of the package correction procedure becomes clear. The temperature profiles obtained by the stack FTM, indeed, highly overestimate the real temperatures, especially in the corners of the dies. The application of the correction methodology allows to account for the high spreading experienced in the corners of the active regions and the estimated temperature is reduced accordingly.

The percentage relative error between the FTM and the corresponding FEM, as well as between the FTM and the measurements, is reported in Figure 7.8 for both the top and the bottom die. From all these graphs it is possible to conclude that the agreement of the FTM is very good, both with the FEM model and with the measurements. This means that the FTM is able to correctly and accurately predict the temperature in real case scenarios. The maximum percentage error is, indeed, less than 1.5% with respect to FEM and less than 4% with respect to experimental results. Moreover, if the average %error all over the die is considered, it is around 1%-1.2% with respect to both FEM and measurements. Table 7.3 lists the precise numbers for the complete FTM (first and second rows) and for the stack FTM (third and forth rows). The reported numbers clearly show the positive impact of the application of the package correction procedure. The maximum error, in particular, is reduced by more than five time with respect to FEM and it is more than halved when compared to measurements.

Due to the natural convection cooling and to the high thermal resistance of the plastic socket, the whole die stack is significantly heated up: the minimum temperature across the die is, indeed, approximately 63°C, which is 38°C above ambient temperature. On top of it, the temperature variation *within* the considered surfaces of the dies is less than 8°C, just 20% of the overall heating up. The high value of the temperature increase in each position of the die helps the %error to be small. Moreover, in these conditions, the package thermal impact is highly significant and the results in Figure 7.7 together with the low value of the %error prove the applicability of the package correction approach.

Looking more carefully to the temperature cross sections and to the error plots, it

Table 7.3: Maximum and average %error in the validation of the FTM and of the stack FTM with respect to FEM and with respect to measurements, in case of LP configuration.

	max(%*err*) top	max(%*err*) bottom	avg(%*err*) top	avg(%*err*) bottom
FTM vs FEM	1.39	1.45	0.84	0.84
FTM vs Meas.	4.04	4.03	1.07	1.23
Stack FTM vs FEM	7.07	7.61	1.66	1.69
Stack FTM vs Meas.	8.13	9.83	2.2	2.21

is evident that the measured temperature is higher than the modeling results on the left-hand side of the graph, while it is lower on the right hand side. This is probably due to some non-ideal conditions in the measurements setup. The PTCQ package is, indeed, placed in a plastic socket and the socket cover is manually screwed to the socket itself to allow good electrical connections between the PTCQ package and the connections of the socket. No thermal grease is used and it might be that some particles remained captured between the socket cover and the PTCQ package and/or that the top surface of the mold compound was not perfectly flat. The package might, therefore, be tilted in the socket and, due to this non-planarity, a higher thermal resistance might have occurred on one side. Due to the high resolution of the temperature maps, this effect is clearly visible in the graphs concerning the cross-sections. It is, nevertheless, a very small impact: the error remains, indeed, always below 4%.

Another interesting comment can be made on the error with respect to the FEM. The FTM is related to the FEM result through the computation of the HSRs and of the package correction profile: if a certain error is present in the FEM model, it is transferred to the FTM. This is why the accuracy of the FTM with respect to FEM is higher than with respect to measurements. The higher value of the error, in this case, is reported in the center of the stack. This is mainly because, in this power map, no power is dissipated in this location. The package correction, however, assumes uniform power dissipation and the values of the heat transfer coefficients are computed in such a way that, for uniform power dissipation, the FTM and the FEM temperatures coincide in the center. For a non-uniform PM, with no power dissipated in the center and a high external thermal resistance, a larger spreading occurs in the silicon than in case of uniform power dissipation. This is due to the better thermal conductivity of silicon with respect to the package material. As a result, since the package correction in the FTM doesn't account for this extra spreading, which causes an extra temperature increase in the center of the silicon, the temperature computed by the FTM in the center is slightly lower than the one computed by FEM. On the other hand, the low value of the error in the corners, confirms, once more, the validity of the package correction approach.

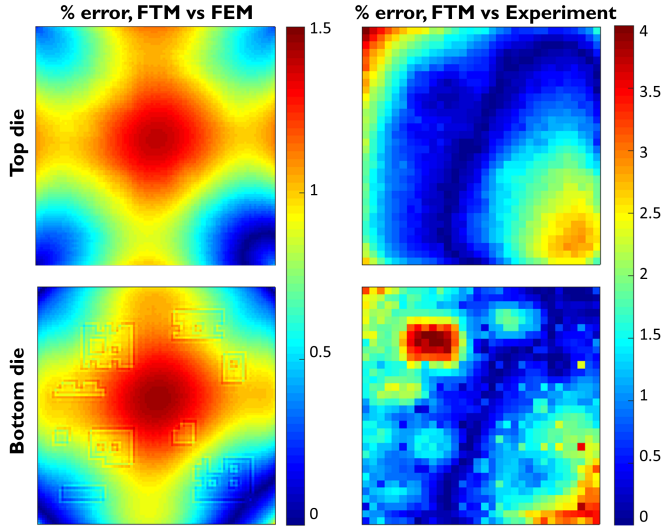


Figure 7.8: %Error of the FTM in the validation of the LP configuration of the PTCQ-on-PTCQ stack. The %error with respect to a detailed FEM is shown on the first column and with respect to the measurements on the second one.

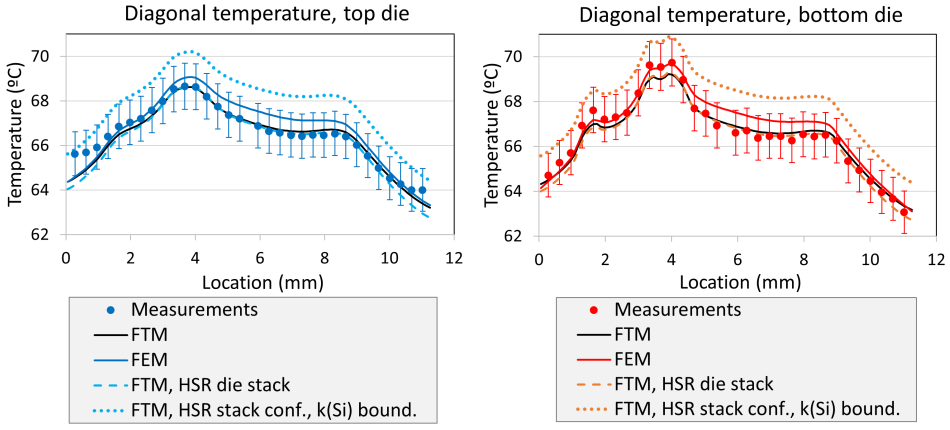


Figure 7.9: Comparison of the temperature profiles obtained by using different k_0 values and different geometries to compute the HSRs for the LP, PTCQ-on-PTCQ test case.

Assessment of a proper selection of k_0 in the LP configuration

Up to now, we discussed the results obtained following the procedure described in Part II of this thesis. The HSRs were computed starting from the stack configuration

Table 7.4: Values of the heat transfer coefficients and of k_0 used in the modeling of the PTCQ test chip (LP configuration) in case different options for the HSRs and for k_0 are considered.

	h_t (W/m^2K)	h_b (W/m^2K)	k_0 (W/mK)
FTM, HSRs stack conf., k_{Si} bound.	1631.15	672.68	136
FTM, HSRs die stack	392.11	392.05	120

(including parts of the package) and the k_0 value assigned to silicon was chosen considering the average temperature increase on the boundaries of the die stack. In this Paragraph, two other options are discussed. The first one consists in computing the HSRs again for the stack configuration but the value of k_0 is calculated considering the average temperature increase on the *boundary surfaces*. In the second case, instead, the HSRs are computed considering just the *die stack* and k_0 refers to the top and bottom surfaces of the stack (they correspond to the boundary surfaces). The values for the corresponding heat transfer coefficients and for k_0 are reported in Table 7.4. The results of this Paragraph highlight, once again, the importance of a proper selection of k_0 and of the HSRs.

Let's consider the first alternative situation. Due to the large amount of low thermally conductive material on top and bottom of the die stack, the computed value of k_0 is strongly different from the one computed considering the average temperature increase on the surfaces of the die stack (Table 7.4). The temperatures, obtained by this first erroneous approach, on the other diagonal of the top and bottom dies (*other* with respect to Figure 7.7), are reported in Figure 7.9 by dotted lines. As it is clear, the error increases drastically: it is more than two and a half time higher than the one computed applying the algorithm presented in Part II of this thesis. The precise numbers concerning the maximum and average percentage errors in this case are reported on the first row of Table 7.5 and can be compared with the ones obtained for the algorithm developed in this thesis and reported in Table 7.3.

The second option is to consider *just* the die stack while computing the HSRs. The package correction is still applied on top of the results obtained by convolving these HSRs with the PMs. The corresponding results are reported in Figure 7.9 by the dashed light blue and orange curves. Following this approach, the computed average temperature on the top and bottom boundary of the die stack (which consists in the layers included between the bottom BEOL and the top die) are, respectively, $T_{0,t}$ = 67.29°C and $T_{0,b}$ = 67.3°C. These values are approximately the same as the computed average temperature increase in the silicon in the original case. As a consequence, $k_0 = 120W/mK$ is assigned to the silicon dies in the computation of the HSRs. The results obtained by this method are very similar to the original ones, which consider the whole stack configuration while computing the HSRs and $k_0 = 120W/mK$. However, it is possible to note that the thermal

Table 7.5: Maximum and average %error obtained by comparing the FTM results, obtained considering different options to compute k_0 and different HSRs structures, with FEM results; LP configuration.

	max(%*err*) top	max(%*err*) bottom	avg(%*err*) top	avg(%*err*) bottom
FTM (HSR package, k_{Si} boundaries) vs FEM	3.59	4.1	2.72	2.74
FTM (HSR die stack) vs FEM	1.83	1.68	1.13	1.11

impact of the package is better included in the former approach. This is evident especially in the corners of the configuration, where the approximation achieved by considering the overmold and the substrate in the HSRs provides better accuracy. Precise numbers concerning the maximum and the average %errors are reported in the last row of Table 7.5.

As a conclusion of this first steady state validation, it is possible to state that the FTM is able to accurately predict the temperature in a LP configuration. The HSRs can be either computed considering the stack configuration or just the die stack, provided that appropriate heat transfer coefficients are applied. The value k_0 of the thermal conductivity to be assigned to the silicon in the calculation of the HSRs must be the one corresponding to the average temperature increase *in* the silicon dies. The inclusion of part of the package in the HSRs (stack configuration) increases the accuracy because the package thermal impact is better accounted for.

High power configuration

In case of the HP configuration, the die stack is not overmolded and an efficient cooling solution is applied on top of the stack (forced convection). For these reasons, the heat is mainly removed from the top and the heat spreading due to the package is limited ($sp = 0.002\%$ according to eq. (5.10)). As a consequence, the HSRs are computed considering just the die stack; the substrate and the Cu-pillars are not included. After calibration of the FEM with respect to measurements in case of uniform power dissipation, the values of the heat transfer coefficient to be applied on top and bottom of the stack are, respectively, defined as $h_t = 14617.6 W/m^2K$ and $h_b = 646.99 W/m^2K$. In this setup, the calculated average temperature, based on the dissipated power map, on the top and bottom surfaces of the stack are, respectively, $T_{0,t} = 43.4°C$ and $T_{0,b} = 45.05°C$. Since only the die stack is considered in the computation of the HSRs and since these values are close to each other, the k_0 value corresponding to their average, $k_0 = 135 W/mK$, is assigned to the silicon material.

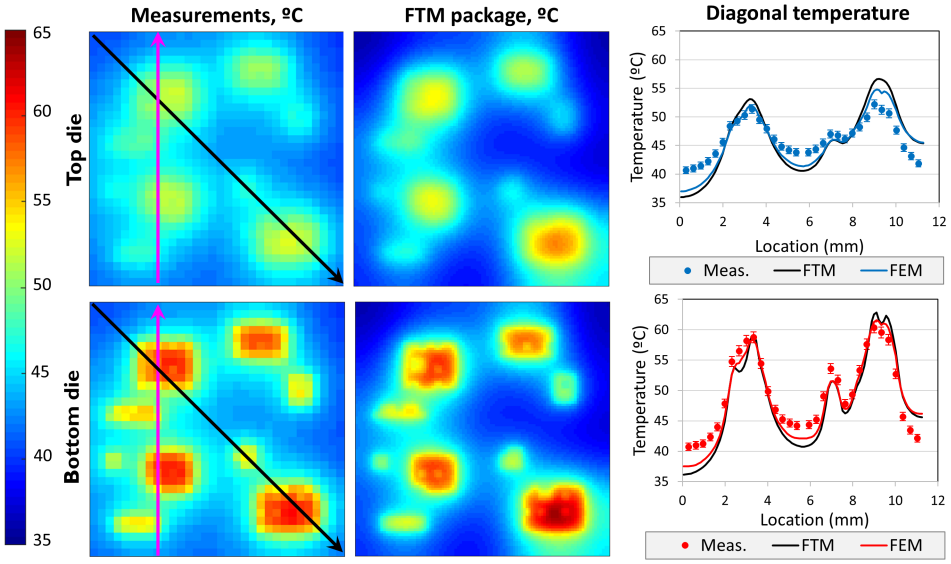


Figure 7.10: Temperature results obtained by measurements and by the FTM for the HP, PTCQ-on-PTCQ configuration. The line graphs on the third column refer to the diagonal cross sections indicated by the black arrows.

For a sketch of the considered geometry, please refer to the details concerning the high power configuration in Figure 5.13 while, for more information about the corresponding FEM model, to Appendix A.1.

The temperature results concerning the HP configuration are shown in Figure 7.10. The first and the second row report, respectively, the data concerning the top and the bottom die. The first column shows the temperature maps obtained by the measurements while the second one the data computed by the FTM (T_{amb} = 25°C, total power Q = 15.3W). The temperature cross sections, along the diagonal indicated by the black arrows, are reported in the third column for both the top and the bottom die. This allows an easier comparison of the obtained results: the measurement data are indicated by colored, circular markers, the FTM by the black, full lines and the corresponding detailed FEM results by colored, full lines. Due to the small impact of the package correction in this configuration, the results of the stack FTM are not reported.

A difference with respect to the LP case that can immediately be noted is the presence of more pronounced temperature peaks. The imprint of the dissipated power map is much clearer in both the top and the bottom temperature maps. This is due to the good cooling applied on top of the device and to the small value of the package spreading resistance, which makes the heat mainly flow vertically. More precisely, the external thermal resistance is responsible for a heating up of the

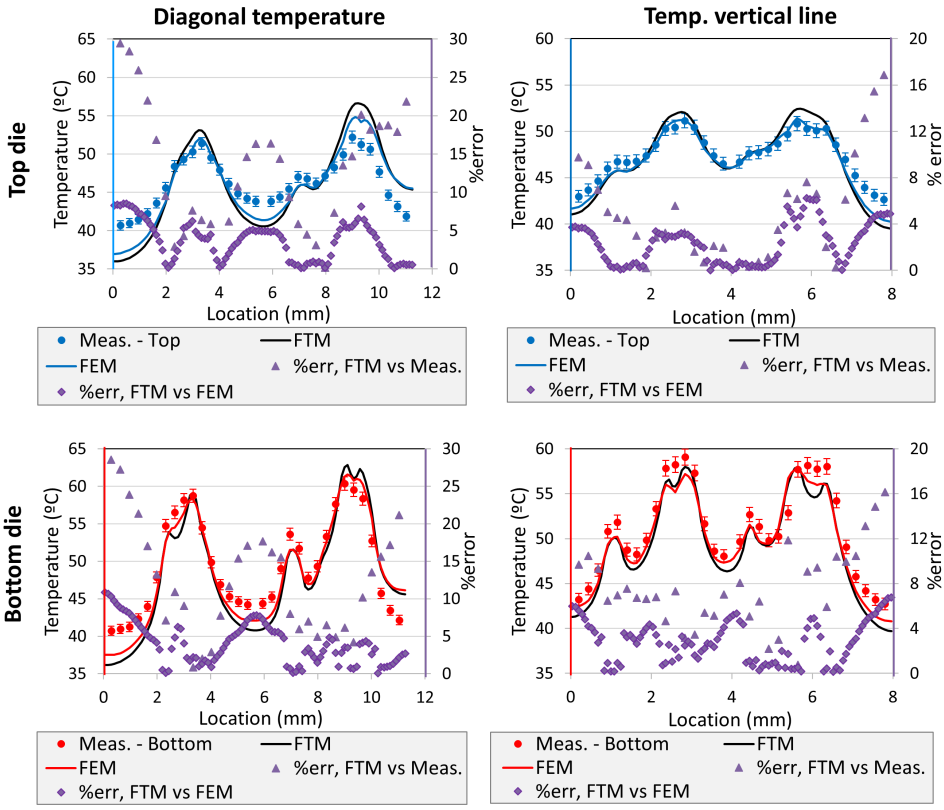


Figure 7.11: Diagonal (first column) and vertical (second column) cross sections along the lines indicated in Figure 7.10. Measured temperature, FEM and FTM results refer to the left axes, the errors to the right axes. Top die results are reported on the first row and bottom die results on the second row.

whole die stack to 37°C (12°C over ambient temperature compared to the 38°C of the LP configuration). On top of this, the localized power causes the temperature value to increase up to more than 63°C. This means that the variation *within* the die (26°C) is much more pronounced than in the LP case.

From the graphs, we can see that both the FTM and the FEM provide a proper estimation of the temperature in the dies. Figure 7.11 reports the diagonal (first column) and the vertical (second column) temperature cross sections along, respectively, the black and the pink arrows indicated in Figure 7.10. The results concerning the top die are reported on the first row, while the ones concerning the bottom die on the bottom row. The red/blue circular markers, the black full lines and red/blue full lines refer to the left vertical axes and correspond, respectively, to the measurements data, the FTM results and the FEM results. The purple markers,

Table 7.6: Maximum and average %error of the FTM with respect to FEM and with respect to measurements. Maximum and average absolute error of the FTM with respect to FEM and with respect to measurements, HP configuration.

	max(%*err*) top	max(%*err*) bottom	avg(%*err*) top	avg(%*err*) bottom
FTM vs FEM	8.85	10.83	3.23	3.93
FTM vs Meas.	29.44	28.52	10.22	11.03
	max |*err*| top (°C)	max |*err*| bottom (°C)	avg |*err*| top (°C)	avg |*err*| bottom (°C)
FTM vs FEM	2.57	1.87	0.57	0.72
FTM vs Meas.	5.28	4.58	1.96	2.26

instead, are used to represent the percentage errors and refer to the right vertical axes: the triangles indicate the %error between the FTM and the measurements, while the diamonds indicate the %error between the FTM and the FEM.

As a first comment, it is possible to state that the errors between the measurement data and the models are higher close to the corners of the stack (edges of the diagonals) and where the power is low. A careful look to the diagonal results reveals the same issue noticed in the LP configuration: the PTCQ package was probably tilted in the socket and, as a consequence, a higher thermal resistance occurred on one side of the package. For this HP scenario, however, the temperature in the corners is much lower than in the LP configuration and, as a consequence, the relative percentage error becomes much higher (up to 30% even). The number concerning the maximum and the average %errors with respect to FEM and measurements are reported in Table 7.6. However, due to the low temperature increases experienced in some locations, the %error could be boosted up due to the low value of the normalization factor. For this reason, the absolute values of the errors are also reported.

Moreover, since both in the LP and in the HP cases, the dissipated power map is the same, if, instead of normalizing the error with respect to the temperature increase, we normalize it with respect to the temperature difference *within* the die, i.e.

$$\frac{|T_{FTM} - T_{ref}|}{\max(T_{ref}) - \min(T_{ref})}$$

where T_{ref} is either the measured or the FEM temperature, the maximum error in the corners is, in both cases around 13%-14%. This means that the FTM approximates equally good the temperature in both the LP and in the HP configurations.

The accuracy of the FTM in the locations of power generation is much higher. If the cross-section along the vertical line (second column in Figure 7.11) is considered,

for example, the maximum percentage error in the locations of high temperature is around 5-6% when the FTM is compared with FEM and around 8% when the FTM is compared with measurements. Since this cross-section is vertical, these experimental results are less affected by the tilt issue. As a consequence, the computed error is lower and more related to the limitation of the FTM methodology itself. Part of this error, in particular, is due to simplifications in the FTM: the interface layer is, indeed, considered as a full μbump array. Since most of the power dissipated on the bottom die is removed from the top, the heat flow has to go *through* the interface layer. The thermal impact of the lack of μbumps in specific locations of the real device is, therefore, quite relevant and causes part of the error. This is, for instance, the reason for the valleys that are visible, in the black curves referring to the FTM results, in correspondence of the peaks of the temperature profiles on the diagonal of the bottom die (around $3mm$ and $9mm$).

The overall accuracy, is however, good enough and this proves the applicability of the FTM to predict the temperature increases also in case of HP steady state configurations.

7.4.3 Transient regime

Data processing

Due to the low value of the dissipated power and to the measurement setup, the experimental data are affected by high noise level. For this reason, the data have been processed to allow for a meaningful validation of the FTM. An example of the processing of the raw experimental data is illustrated in Figure 7.12. The typical measurement of one diode for a time of n seconds consists in a heating up and in a cooling down phase. The time length of the heating up phase is approximately $n/2$ while, since the moment in which the heaters are powered might differ from zero, the cooling down phase might be shorter. As already mentioned, the temporal resolution of these measurements is $n/200000$. In each experiment, the voltage both in the diode and in the heaters is recorded at each time step. The knowledge of this last set of data, together with the one of the current provided to the system, allows the computation of the overall dissipated power. Due to the limitations of the setup and of the PTCQ test chip, the total amount of dissipated power is low and varies between 0.6W and 0.7W.

From the data concerning the voltage in the heaters, moreover, the exact moment in which the power starts to be supplied and in which it ceased to be supplied can be derived. This allows to split the data into two parts: the *heating up* and the *cooling down* phases (cf. Figure 7.12 (b)). In order to reduce the noise, a two-steps procedure is performed. Zooming-in into the original set of data (cf. Figure 7.12 (a)), it is possible to note that the noise can be split into a lower and a higher frequency component. The low frequency component is, in particular, related to

the 50 Hz frequency of the alternated current. In order to filter out this noise, a *moving average filter* with a span of $span = \frac{1}{50}\frac{200000}{n} + 1$ is implemented. This results in the red lines (below the black dots) in Figure 7.12 (b).

Now that the smoothed curves are obtained, the next step consists in reducing the dimension of the data vectors: the ones obtained after the smoothing have still 200000 elements each, uniformly separately in linear scale. Due the huge difference in the time scale at which the different materials in the package respond to the dissipated power from a thermal point of view, the transient thermal analyses are normally performed in logarithmic time scale. The smoothed experimental data are, therefore, sampled accordingly: a higher rate of data is kept close to the beginning of the process and more data are discarded close to the end of the process. More precisely, the data that are kept for the validation of the FTM are the ones that occupy positions

$$\mathrm{cumsum}\left[\mathrm{round}\left(\exp^{\log(a)\cdot\frac{[1:b]}{b}}\right)\right]$$

in the vector containing the smoothed measurement data. $[1:b]$ indicates a vector of values from 1 to b, spaced 1 while a and b are natural numbers and indicate, respectively, how many points are approximately skipped towards the end and the number of elements that the reduced data vector contains. This basically means that the points to be kept are chosen so that the distances between them is approximately uniformly distributed in logarithmic scale. The term *approximately* is used because of the *round* command that rounds the numbers to integer values. The values $a = 5000$ and $b = 250$ are used in the following validations. The obtained values are indicated with black dots in Figure 7.12 (b).

The last step in this procedure consists in getting the temperature increase (or decrease) profile from the variation of the measured voltage. This is obtained by dividing the voltage increase/drop by the sensitivity of the diode ($\sigma = -1.55 mV/°C$). The results are reported in Figure 7.12 (c). The cooling down curve starts after the end of the heating up phase but, to take advantage of the representation in the logarithmic time scale, its start-time is always shifted to zero.

Modeling information

The experimental validation of the FTM in the transient regime concerns the LP configuration. Two cases will be analyzed in the following of this Section, corresponding to the PMs shown in Figure 7.13. Both cases consist of two dissipated HS, one on the top and one on the bottom die, of 15 cells each. These active cells are arranged in 4×4 arrays including one cell without power dissipation (cell not of type #2). The two HSs considered in each case are aligned near the center of the die for *Case1* and in the corner of the die for *Case2*. In both cases the power map dissipated on the top die is equal to the one dissipated on the bottom die. Four different diodes are measured in each case: two are located approximately in the

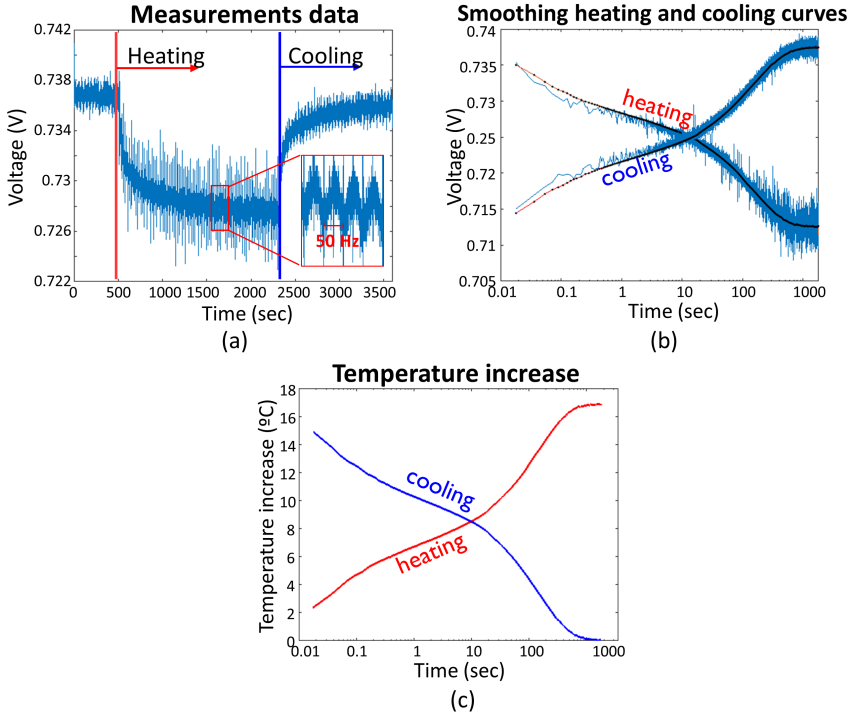


Figure 7.12: Processing of the transient experimental data to obtain the temperature curves for the model validation.

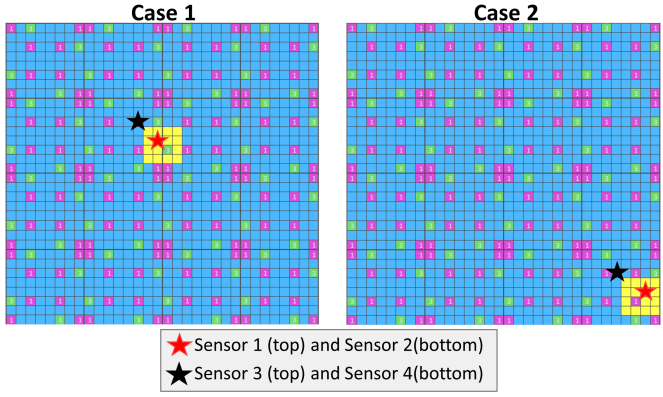


Figure 7.13: Power maps for the two cases analyzed in the transient validation of the PTCQ-on-PTCQ stack.

Table 7.7: Values of the heat transfer coefficients used in the modeling of the packaged PTCQ test chip in the transient experimental validation, LP configuration.

$h_{t,pack}$ (W/m^2K)	$h_{b,pack}$ (W/m^2K)
462	238

center of the HSs (red stars in Figure 7.13) and two in the corner, just outside the HSs (black star in Figure 7.13). More precisely,

Sensor 1: top die, near the center of the HS;

Sensor 2: bottom die, near the center of the HS;

Sensor 3: top die, outside corner of the HS;

Sensor 4: bottom die, outside corner of the HS.

Due to the importance of the capacitance values of the different parts of the system and, in particular, of the mold compound, of the substrate and of the PCB, the whole stack configuration (which includes these parts) is considered in the computation of the HSRs. Moreover, as already mentioned in the beginning of this Section, the substrate material is divided into two parts. This is because the FEM, developed for the steady state configuration, proved to lack the capacitive impact of the PCB. The trick of dividing the substrate into two parts allows to re-use the same modeled geometry and, at the same time, to account for the capacitive effect of the PCB.

Another difference with respect to the steady state model is in the applied BCs. This is probably due to some differences in the measurement setup and environment. More precisely, the BCs for the detailed package FEM have been defined during the calibration process (comparing the measurement data and the FEM results from 100μsec to 30 minutes). The heat transfer coefficients applied on top and bottom of the package configurations in the FEM and FTM model are reported in Table 7.7, while the modeled geometry is sketched in Figure 5.13 (LP configuration). The heat transfer coefficients to be applied on top and bottom of the 2D-model used to calculate the HSRs have, then, been derived so that, for uniform power dissipation, the maximum temperature in the package and in the stack configurations are the same, once the steady state have been reached ($h_t = 5589W/m^2K$ and $h_b = 933.68W/m^2K$).

Concerning the value of the silicon thermal conductivity to be used in the HSRs, the value of $k = 148W/mK$ has been chosen. This is because, due to the limitation of the current and voltage suppliers, the total dissipated power varies between 0.6W and 0.7W. This means that the expected average temperature variation in the silicon is around 1-2°C. Moreover, the Kirchhoff transformation won't be applied

in the following transient experiments because, for this amount of dissipated power, the maximum temperature increase is not expected to be high enough to make it relevant. As shown in Chapter 6, a high temperature difference within the die is needed for the Kirchhoff transformation to effectively improve the FTM results. Moreover, as already explained in Subsection 7.4.2, this LP configuration has a high external resistance that efficiently smears out the temperature peaks. It should be noted that the normalized temperature increases are reported in the following graphs: this is because the dissipated power might slightly change from one measurement to the other. Normalizing the results allows a better comparison.

Another remark concerns the package correction. As illustrated in Chapter 5, for the transient regime, the equivalent material properties of the die stack need to be computed and used in the coarse FEM of the package. The computed values are, in particular, $k_{xy} = 139W/mK$, $k_z = 43.5W/mK$, $c = 754.52J/kgK$ and $\rho = 2330kg/m^3$.

Experimental validation

Figure 7.14 shows a first set of results. The graphs in the first column concern *Case1* (HS in the center) while the ones in the second column refer to *Case2* (HS in the corner). The location of the sensors is indicated in Figure 7.13 and it is sketched also in Figure 7.14. Red curves indicate the processed measurement data, green curves refer to the FEM results, orange circles to the FTM including the package correction and blue segments to the FTM without package correction. The reported data are the normalized temperature increases and they are shown in logarithmic time scale. The first thing we can notice is that the agreement between all the curves referring to the same situation is good. The maximum and the average error over time in each case are reported in Table 7.8.

The good agreement between the measurement data and the FEM proves, in particular, that the FEM has been successfully calibrated. The correct definition of the material properties and of the heat transfer coefficients is, indeed, a crucial step in order to be able to obtain accurate results with the FTM. Concerning the FTM, a **variable time step approach** has been implemented and, as a consequence, the dots in the plots are not equidistant. If the transient FTM with constant time step, which has been presented in the previous Chapters, would have been adopted here, 10000 time steps would have been needed to obtain the temperature values in the range 100μs-1sec. Fortunately, under certain conditions, the FTM can be implemented with a variable time step. In order to achieve this aim, the HSRs and the package correction profiles have to be stored at the desired, non-constant time steps. In Figure 7.14, for example, 20 linearly distributed points per decade have been considered. It is important to stress that the length of the time steps has to be defined by the user: this is not an automatic procedure. Moreover, since the whole FTM is based on superposition, this approach can only be applied if the dissipated PM is constant in time or if the variation of the time step is selected in a proper way. In this latter case, in particular, the lengths of the time steps at which the

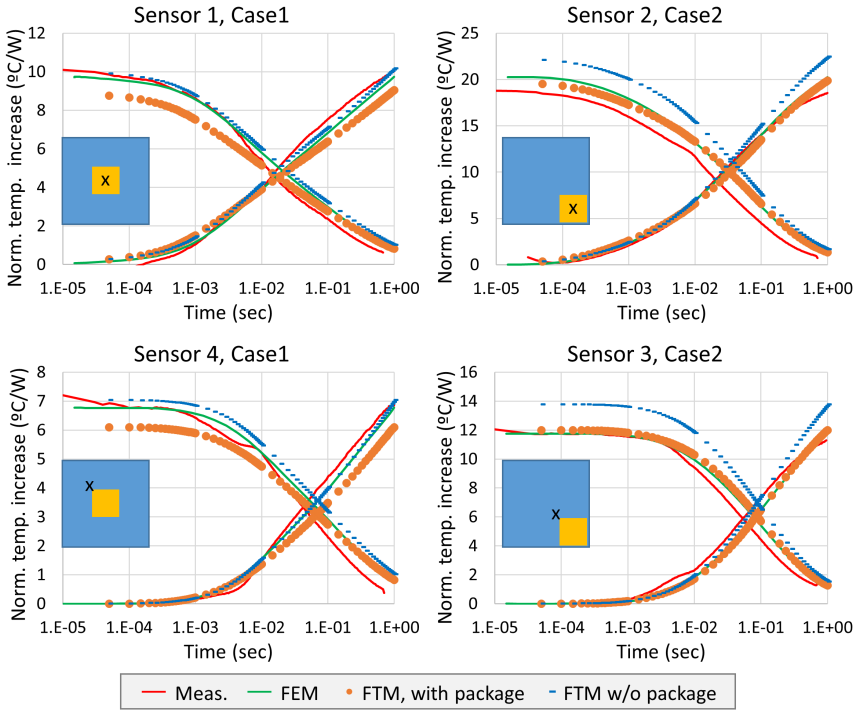


Figure 7.14: Short time scale transient validation of the PTCQ-on-PTCQ test chip. Case1: HS in center; Case2: HS in corner; Sensor1&Sensor2: diode aligned with HS; Sensor3&Sensor4: diode outside the HS.

HSRs and the package correction profiles are stored need to be *periodically repeated*, with a period equal to the greatest common divisor of the time during which the power map does not change. If, for instance, a time sequence of PMs as

$$PM_1,\ PM_1,\ 0,\ 0,\ 0,\ 0,\ PM_2,\ PM_2,\ PM_2,\ PM_2,\ PM_3,\ PM_3,$$

is considered, the time discretization of the HSRs can be variable in the interval $(0, 2\Delta t]$. The same discretization needs, then, to be repeated in all intervals $(2n\Delta t, 2(n+1)\Delta t]$, $n \in \mathbb{N}$ until steady state is reached. It is important to note that, for this algorithm to work, the units of the PM need to be in W while the ones of the HSRs in $°C/J$ (PMs divided by time steps). The temperature is, then, computed as

$$T_{z_i}(\bar{i}, \bar{j}, z_j, \bar{t}_k) \approx \sum_{\bar{t}_l=1}^{\bar{t}_k} HSR_{z_i}(\cdot, \cdot, z_j, \bar{t}_l) *_{2D} PM_{z_i}(\cdot, \cdot, \bar{t}_k - \bar{t}_l)\Delta\bar{t}_l + T_{amb}. \tag{7.1}$$

By multiplying the intermediate results by the time step corresponding to the considered step in the HSR, the same time variation is considered for PMs and

Table 7.8: Maximum and average absolute error of the FTM (with and without package correction) with respect to measurements (with respect to FEM in brackets) for the four cases in Figure 7.14.

	Sensor 1, Case1	Sensor 4, Case1	Sensor 2, Case2	Sensor 3, Case2
max $\|err\|$, FTM w. package ($°C/W$)	1.32 (1.07)	1 (0.68)	2.33 (0.74)	1.34 (0.39)
avg $\|err\|$, FTM w. package ($°C/W$)	0.66 (0.56)	0.46 (0.39)	1.6 (0.32)	0.69 (0.23)
max $\|err\|$, FTM w/o package ($°C/W$)	0.92 (0.43)	0.9 (0.27)	4.22 (2.16)	2.42 (2.02)
avg $\|err\|$, FTM w/o package ($°C/W$)	0.49 (0.31)	0.54 (0.22)	3.36 (1.5)	1.98 (1.42)

HSRs that are convolved together. This small variation of the FTM algorithm allows to follow the transient thermal behavior of the system at different time scales. This was not possible by using constant time step unless computing the temperature in a huge amount of points, which might make the FEM with adaptive time step computationally more convenient than the FTM. However, if the temperature is needed only at certain constant time steps, without more precise information concerning the heating up/cooling down phases, the algorithm with constant time step (Δt or $2\Delta t$ in this case) is preferable and faster.

A closer look to the error (cf. Table 7.8 and Figure 7.14) reveals that, for *Case1* (HS in the center), the error is almost always below $1°C/W$. Just in case of the sensor in the center of the HS on the top die it is a bit higher. This worse result occurs in case the package correction is included. As both the Table and the Figure show, indeed, in case the HS is dissipated in the center of the stack, the application of the package correction slightly worsens the results. This is because, as already mentioned for the steady state validation of the LP configuration, the correction profiles have been calculated assuming uniform power dissipation. Due to the high thermal resistance of the overmold material, if the HS is dissipated in the center of the stack, the heat spreads into the silicon more than in case of uniform power dissipation. Moreover, for a HS power dissipation *in the center* of the stack, the impact of the thermal spreading due to the package is limited. It is, in particular, less significant than the extra spreading in the silicon. Nevertheless, the accuracy of the results obtained applying the package correction is still acceptable.

For *Case2* (HS in the corner), instead, the absolute errors are higher. This is mainly due to the higher temperature values reached by the system in this position. For this scenario, however, the application of the package correction highly improves

the accuracy of the results. In this location, indeed, the impact on the temperature increase of the thermal spreading due to the package is much more significant than the extra impact of the spreading in the silicon due to HS instead of uniform power dissipation. The inclusion of the package thermal impact allows, in particular, to reduce the error by half.

Figure 7.15 reports the results referring to the same cases as in Figure 7.14 but for a **longer time scale**. Now, both the heating up and the cooling down phases last 5 seconds instead of 1 second. For this reason, the results are shown in linear time scale and the FTM is run with a constant time step. The legend of this new Figure is the same as the one used in the short-time analysis, except for the markers used for the FTM with package correction, which are, now, blue triangles. Analogous comments over the accuracy of the FTM as the ones made for the short time analysis are appropriate here. The plots in linear time scale highlight, in particular, the importance of the package correction: its application reduces the error of approximately 5°C/W if the HS is dissipated in the corner. On the other hand, for HS in the center, the overcompensation of the FTM results due to the application of the package correction, is not so drastic. This means that, in case of hesitation whether to include or not the package thermal impact in the FTM, its inclusion when it's not necessary, is much better than not including it when it's needed.

Up to now the transient validation has been performed considering just one heating up and the corresponding cooling down phase. In Figure 7.16 two scenarios with multiple on-off switchings of the power dissipation are analyzed for the LP configuration. The thermal behavior of the system is followed for 20 sec in case of a 90% (blue) and a 10% (red) **duty cycle**. Five pulses have been considered in each case. The full lines in the Figure refer to the normalized measurements results while the circular markers to the FTM results obtained including the package correction. The green line represents the normalized temperature increase in case the heaters are maintained continuously active for 20 sec. For this test case, the power is dissipated according to the power map of *Case1* (HS in the center) but just in the bottom die, no power is dissipated on the top die. The data collected from *Sensor 1* (top die, inside the HS) are considered.

The agreement between the measurements and the FTM results is very good for both analyzed cases. The maximum error in case of 90% duty cycle is $2.37°C/W$ while the one in case of 10% duty cycle is $1.11°C/W$. The corresponding average errors are, respectively, $0.17°C/W$ and $0.13°C/W$. These results concerning the duty cycles confirm that the FTM model is able to account for both the short time scale and the long time scale heating effects. This is, indeed, highlighted by the fact that the normalized temperature increase obtained from short pulses, which account mainly for the short-time scale heating, always returns to approximately zero, while the one obtained for the 90% duty cycle, which includes also the impact of the package, increases with time. In both cases the accuracy is very good.

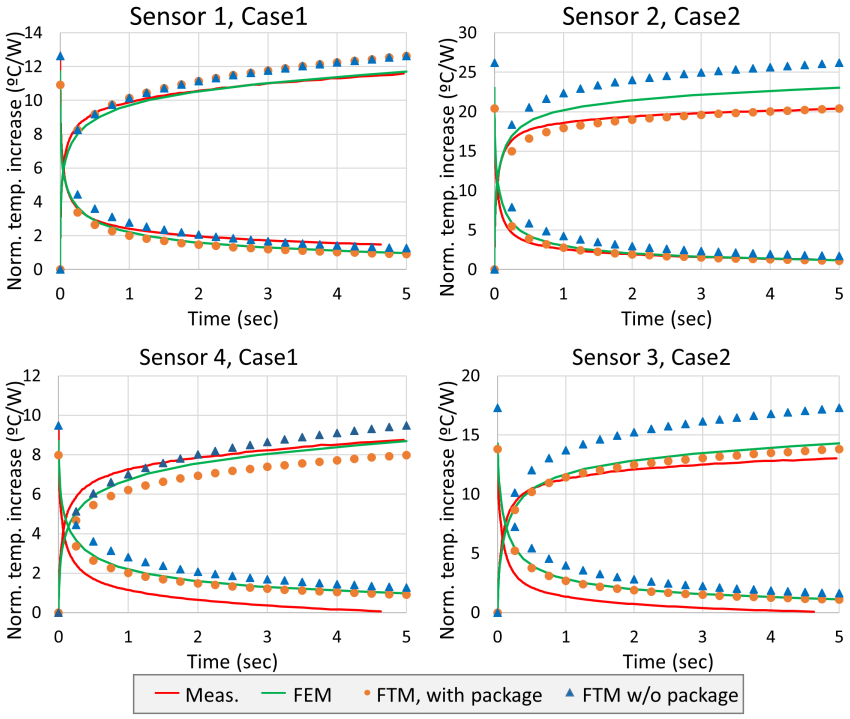


Figure 7.15: Longer time scale transient validation of the PTCQ-on-PTCQ test chip. Case1: HS in center; Case2: HS in corner; Sensor1&Sensor2: diode aligned with HS; Sensor3&Sensor4: diode outside the HS.

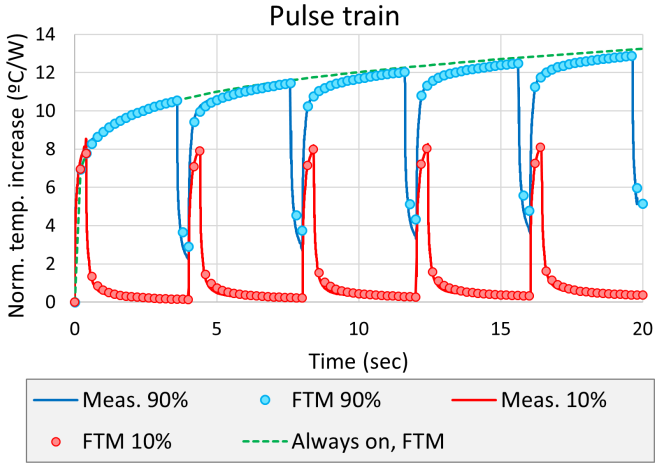


Figure 7.16: Transient PTCQ validation, pulse trains: 90% duty cycle in blue and 10% duty cycle in red.

7.5 Summary

In this Chapter the FTM has been successfully validated with respect to measurement results in both the steady state and the transient regime for a packaged, two dies, 3D-IC stack. Two configurations have been considered for the steady state regime: a low power configuration, with high external thermal resistance, and a high power configuration, with a heat sink and forced convection cooling. In both cases, a non-uniform power map has been dissipated and the full temperature maps have been calculated. The accuracy of the FTM results with respect to the measurements proved to be good all over the dies.

For the transient regime, only the LP configuration has been analyzed. Due to the limitations of the test vehicle and of the measurement setup, HS power dissipation has been considered and the temperature increase has been monitored in isolated locations. The temperature results have been analyzed for both the short and the long time scale. To be able to follow the thermal response of the system for several decades (from 100μsec to 1 sec), the possibility to use a variable time step in the FTM has been presented, together with the related limitations. Finally, the performance of the FTM for a case study, in which different duty cycles have been considered for 20 seconds of chip activity, has been analyzed showing very good accuracy. All these results prove that the transient FTM is able to account for both the localized heating (short time) and the general heating of the whole system (long time). The results are, indeed, obtained with high accuracy with respect to both FEM and experimental measurements.

Chapter 8

Extensions of the Methodology to Different Geometries

8.1 Introduction

The FTM presented in this thesis has been developed to estimate the temperature increase in case of packaged 3D-ICs in which multiple dies, all with the same size, are stacked on top of each other. In this Chapter, the methodology is extended to deal with different geometries commonly available for microelectronic applications. Two cases are considered: an *interposer configuration*, in Section 8.2, in which the active dies are places next to each other on top of a common interposer, and a *pyramidal configuration*, in Section 8.3, in which the dies are still stacked on top of each other but they have different dimensions (cf. Figure 8.1). In both cases, a larger structure (interposer or larger die) is present at the bottom of the geometry. The thermal impact of this structure can be comparable with the one of a conventional package for the standard 3D-stack: its larger area with respect to the active regions allows thermal spreading to occur. For this reason, similar methodologies, as the one presented to include the package thermal impact on top of the results obtained for the stack configuration in Chapter 5, are proposed.

The analysis for the interposer is performed in steady state and validated with respect to FEM results. Just the case of a passive interposer (no power dissipation in the interposer itself) is considered in this thesis. Transient experimental data are available for a test chip in which the top die is smaller than the bottom one. To take advantage of these data, the FTM for stacks of dies with different sizes has been developed in the transient regime and validated with respect to both FEM models and experimental results. Since the steady state regime is a special case (final stage) of the transient regime, the steady state results are not reported.

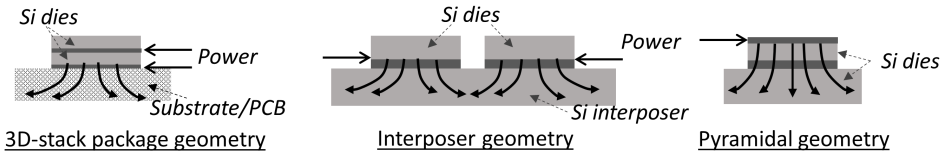


Figure 8.1: Schematic of the geometries of packaged 3D-stack, of the interposer and of the pyramidal configurations. The thermal spreading due to the larger section of the geometry is also sketched.

8.2 Interposer configuration, steady state

In this Section, the FTM methodology is extended to allow the steady state thermal analysis of the silicon interposer configuration. An example of the interposer geometry is sketched in the central part of Figure 8.1. The main characteristic of this configuration is that more active dies are integrated on top of a passive, silicon (laminate or glass), larger die. The number and the location of the dies, placed face down on top of the interposer (F2F configuration), can vary and it is application dependent. All these dies are, in particular, connected to the PCB via TSVs, as in a 3D-stack, going through the interposer die. The interposer itself can be active, performing some functions, or passive. In this thesis, just the passive option is analyzed.

If, on the one hand, this kind of geometry has a larger footprint with respect to a 3D-stack, loosing from a miniaturization point of view, it still uses, on the other hand, the fine pitch, high density connections of the 3D-TSV technology, allowing fast and short connections. Moreover, it performs better from a thermal point of view. This is because the different dies are arranged *next to* each other instead of *on top of* each other. This means that more thermal spreading occurs in the interposer itself and that the area available for cooling does scale with the number of dies. The whole structure can, then, be packaged in several ways depending on the specific application (placing a lid on top of the dies, overmolding them or leaving them exposed).

8.2.1 Modeling methodology

The fundamental idea to extend the FTM methodology to this new configuration is based on the application of an *appropriate* package correction on top of the results obtained by the stack FTM for a layered structure. Due to the specific geometry of the interposer, the stack configuration, on which the stack FTM is applied, can be chosen in two different ways. The layers included in the stack configuration are, in both cases, the ones located in correspondence of the position of the active dies; what changes are the dimensions of the modeled stack itself. In the former case

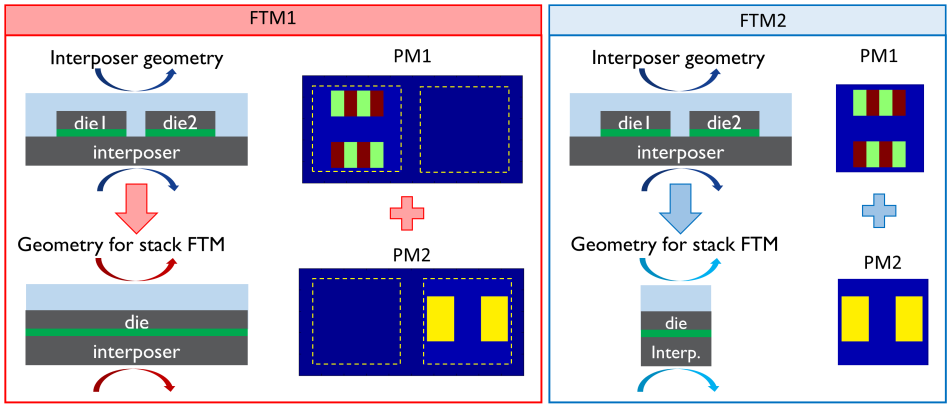


Figure 8.2: Comparison between the two modeling methodologies presented for the interposer configuration.

(referred to as *FTM1*), the specified layered structure is extended to obtain a stack configuration as large as the *interposer* itself while, in the latter case (referred to as *FTM2*), it is considered as large as the *die*. This difference is sketched in Figure 8.2 and more details are given hereafter.

Before further commenting on these two methodologies, it is important to note that, in both cases, the thermal impact of the dissipated power map on each die is considered separately. Superposition is applied afterwards to include the mutual impact of all the dies. Moreover, neither of these options is limited to a fixed number of dies nor to dies all with the same dimension. In case of *FTM2*, in particular, the dimension of each die is accounted for by using the method of images in the stack FTM. This means that dies with different dimensions can be easily handled by this model. The package corrections are, then, computed considering the complete real geometry and they account, therefore, for the exact size of each die.

Stack configuration as large as the interposer, *FTM1*

The algorithm for this first approach, in case of N dies, can be summarized in the following steps.

1. Application of the FTM for the stack configuration. This requires the computation of the basic ingredients needed for the FTM.

   HSR: The HSR (just one in this case) is computed by means of a 2D-axisymmetric, FEM model.

PMs: The N power maps, one for each die, are as large as the considered stack (interposer size). In each of them, just the power dissipated on one single die is considered; the values corresponding to the remaining area are set to zero (cf. Figure 8.2). It is important to stress that power can be dissipated just in the location of the dies: uniform power dissipation on the dies *does not mean* a uniform PM.

N^2 temperature increase profiles, Θ_{ij} on die j due to the specific power map dissipated on die i, are computed by applying the stack FTM methodology to the considered stack geometry.

2. Creation of the coarse FEM for the *interposer geometry* to extract the data needed to compute the correction profiles. There are few small differences with respect to the coarse model developed for the 3D die stacks.

   - All the layers are included in the coarse model, the die stack is not considered as full silicon (cf. Paragraph *"Modeling of the stack in the package configuration"* in Section 5.3.4). This is because the correction profiles are computed considering uniform power dissipation in the locations where power *can* be dissipated. Previously, for uniform power dissipation in the *stack configuration* of the *3D-package*, a uniform temperature profile was obtained. That profile was easily achievable by a simple resistance network model. It was, therefore, easy to compute the difference in temperature increase between a stack configuration composed by multiple layers and one composed by one single material. This difference was, then, added to the results obtained by the simplified FEM model for the package configuration in which the die stack was assumed of full silicon.

     In the stack configuration for the *interposer* geometry, however, uniform power dissipation on one die does not mean uniform power dissipation all over the equivalent stack configuration for the interposer. The PM is not uniform everywhere but uniform power is localized at the dissipating die itself. As a consequence, thermal spreading plays a role in this computation. A simple resistance network can't, therefore, be used to include the thermal impact of layers with different material properties on top of the results obtained considering a simplified coarse FEM model in which the die stack is constituted by a single material. For this reason, all the layers, with their own material properties, have to be included in the coarse FEM model.

   - The temperature increase profiles for the interposer configuration are obtained by cubic spline interpolation of the data extracted from the FEM model and not by scaling an initial, general surface (cf. Paragraph *"Scaling of a basic surface"* in Section 5.3.4). This is because the interposer package is completely different than the 3D one; it is, in particular, not necessarily symmetric. The interpolation approach is also more

convenient in case an interposer geometry with more dies or with dies in different positions has to be considered.

This coarse FEM model is run N times, each time dissipating uniform power in one die and extracting the temperature profiles in the locations corresponding to all the dies in the interposer. This results in N^2 temperature increase profiles named $\Theta_{FEM,unif,ij}$, where index i indicates the die where power is dissipated and index j the one where temperature is computed.

3. Computation of the temperature increase profiles, $\Theta_{FTM1,unif,ij}$, for uniform power dissipation on each single die in the *stack configuration*. In this case, as already mentioned, uniform power dissipation on *one die* doesn't mean uniform power dissipation all over the *stack configuration*. An approach based on a resistance network or on the *annulus method* (cf. Section 5.3.4) can't, therefore, be applied. As a consequence, the stack FTM is applied N times, each time assuming uniform power dissipation only on one die. Every time, the temperature profiles on each of the N dies are extracted, resulting, therefore, in N^2 surfaces.

4. Computation of the N^2 correction profiles, $\bar{C}_{int,ij}$, on die j for the active die i as
$$\bar{C}_{int,ij} = \frac{\Theta_{FEM,unif,ij}}{\Theta_{FTM1,unif,ij}}. \tag{8.1}$$

5. Computation of the final temperature profile on each die as
$$T_j = \sum_{i=1}^{N} \Theta_{ij} \cdot \bar{C}_{int,ij} + T_{amb}. \tag{8.2}$$

Stack configuration as large as the die, *FTM2*

The algorithm for this second approach is quite similar to the first one. The main difference is in the geometry of the stack configuration. More precisely, it consists in the following steps.

1. Application of the FTM for the stack configuration. This requires the computation of the basic ingredients needed for the FTM.

   HSR: The HSR (just one in this case), is computed by means of a 2D-axisymmetric, FEM model. It is the same as for algorithm *FTM1*.

   PMs: The N power maps, one for each die, are as large as the considered stack, which, in this case, has the size of the die itself (cf. Figure 8.2). This methodology is not restricted to dies with the same size because the real dimension of each die is included by means of the method of images.

This step creates N self heating temperature profiles Θ_{ii}, due to the specific power map dissipated on die i and the considered stack configuration.

2. Creation of the coarse FEM for the *interposer geometry* to extract the data needed to compute the correction profiles.

   - Concerning the material considered in the location of the die stack, both the approaches presented for the 3D-package (full Si stack) and for the interposer configuration in the *FTM1* algorithm (layered stack) can be considered. The first approach is possible because, now, when the *stack configuration* is considered, uniform power dissipation on one die does mean uniform power dissipation on the stack configuration. As a consequence, in the stack configuration, the difference in the temperature increase between a model with a full Si die stack and another more realistic one with a layered die stack, can be computed by a resistance network. This difference can, then, be included in the simplified FEM model for the *interposer geometry* in which the die stack is assumed to be made of full Si. However, since the achievable reduction in the number of considered layers in the simplified FEM model is limited (just one die plus BEOL and interface constitute the die stack), meaning that the gain in computational time is also limited, the FEM is run considering all the different layers in the die stack.
   - Cubic spline interpolation is used to extract the temperature increase profiles from the coarse FEM (as for *FTM1*).

   This means that the extraction of the temperature profiles from the coarse FEM for the interposer geometry works in the same way as for the *FTM1* algorithm (point 2 in the numbered lists of the two algorithms is the same). Therefore, also in this case N^2 temperature profiles, $\Theta_{FEM,unif,ij}$, are extracted in the locations of both the active and of the passive dies.

3. Computation of the temperature increase profiles, $\Theta_{FTM2,unif,ii}$, for uniform power dissipation in the stack configuration. Since, in this approach, uniform power dissipation in *one die* does mean uniform power dissipation on the *stack configuration*, either a resistance network approach or the *annulus method*, illustrated in Paragraph *"Temperature for stack configuration"* in Section 5.3.4, can be employed.

4. Computation of the N correction profiles for the active dies as

$$\bar{C}_{int,ii} = \frac{\Theta_{FEM,unif,ii}}{\Theta_{FTM2,unif,ii}}. \tag{8.3}$$

5. Computation of the temperature increase profiles, $\Theta_{int,ij}$, $i \neq j$, on the passive dies. From multiple FEM simulations and experimental results, the temperature increase on the passive dies proved to be mainly related to

the *total amount of power*, Q_i, dissipated on the active die, rather than to the specific power map. For this reason,

$$\Theta_{int,ij} = \Theta_{FEM,unif,ij} \frac{Q_i}{Q_{int,unif,i}}, \qquad i \neq j \tag{8.4}$$

where $Q_{int,unif,i}$ is the total power dissipated on die i while computing $\Theta_{FEM,unif,ij}$.

6. Computation of the final temperature profiles on each die as

$$T_j = \sum_{i \neq j} \Theta_{int,ij} + \Theta_{jj} \cdot \bar{C}_{jj} + T_{amb}. \tag{8.5}$$

Test structure

The test structure, which is used for the validation of the FTM for the interposer geometry, consists of a configuration with two $8 \times 8mm^2$ active dies placed, face down, on top of a $20 \times 10mm^2$ interposer. Figure 8.3 (a) shows the top view (half symmetry) and the lateral cross section (not to scale) of the analyzed case. A lid package is considered and the space around the dies is assumed to be empty. This package configuration, appropriate for a middle-power application, has a copper lid that encapsulates the whole structure. This lid, which is used to improve the thermal spreading, is attached to the top of the dies via a thermal interface material (TIM). In this way, the thermal coupling between the dies happens both through the interposer (below) and through the lid (above). Since the original idea was to experimentally validate the interposer FTM, the measurement environment has been considered in the model. This means, for example, that, based on the measurements set-up illustrated in Chapter 7, a spacer is modeled on top of the lid. However, due to some issues with the measurements procedure, the experimental validation has not been possible and just the FEM validation is reported. Concerning the bottom part of the package, the interposer die is connected to the package substrate via copper pillars. Equivalent material properties are used for the layers in which multiple materials are present (silicon interposer + TSVs, μbumps + underfill, BEOL, copper pillars,...). More details about the dimensions and the material properties of the different layers in the structure are reported in Table 8.1 [71].

The considered test case is, therefore, a *packaged interposer* and, as it will be shown in the following of this Section, the FTM is able to include both the thermal impact of the interposer die itself and of the package. This is possible by applying just one single correction profile because the coarse model, from which the correction profiles are extracted, includes both the interposer die and the package.

In Figure 8.3 (b), the temperature increase profiles on both dies, for uniform power dissipation on the left die, is shown. As expected, the temperature on the active

Table 8.1: Parameters used in the FEM and FTM simulations of the interposer, listed starting from the top of the geometry. The notation (×2), means that the structure is repeated two times.

Layer (from top)	Dimensions ($mm \times mm \times mm$)	k or k_x, k_y, k_z (W/mK)
Spacer	$35 \times 35 \times 2.8$	180
Lid	$28.4 \times 28.4 \times 0.285$	400
TIM (×2)	$8 \times 8 \times 0.095$	2
Silicon dies (×2)	$8 \times 8 \times 0.2$	150
Power dissipation and temperature computation		
BEOL dies (×2)	$8 \times 8 \times 0.002$	$0.25 \times 0.25 \times 0.5$
Interface (×2)	$8 \times 8 \times 0.013$	$0.5 \times 0.5 \times 14$
BEOL interposer	$20 \times 10 \times 0.01$	$0.25 \times 0.25 \times 0.5$
Silicon interposer	$20 \times 10 \times 0.1$	150
Cu pillars	$20 \times 10 \times 0.1$	$0.4 \times 0.4 \times 8$
Substrate	$35 \times 35 \times 1.16$	$12 \times 12 \times 0.6$

die is much higher than the one on the passive die and the peak values of the temperature are not in the centers of the dies. For the active die, in particular, the location of the maximum temperature is slightly shifted toward the left (away from the other die). This is because a larger thermal spreading is experienced on the right hand side of this die due to the presence of more material in the geometry in that direction. For the passive die, instead, the temperature decreases while moving towards the right hand side, because the heat source is on its left.

Comparison between *FTM1* and *FTM2*

Figure 8.4 reports a comparison between the results obtained by applying algorithms *FTM1* and *FTM2* to the illustrated test case. In particular, in plot (a), the comparison between the correction profiles on the diagonal of the dies is presented. A more accurate discussion about this topic is reported later on in this Section but, as a first initial comment, we can clearly see that the correction profiles referring to the two methodologies are significantly different from each other. In particular, if the approach in *FTM1* is selected, the results obtained by the stack FTM are extremely different from the expected ones. With the second approach, instead, the value of the correction profile is around 1, meaning that the application of the FTM to the stack configuration considered in this case already provides a quite accurate temperature estimation. No curve is reported for the passive die in case of algorithm *FTM2* because this profile does not appear in the methodology.

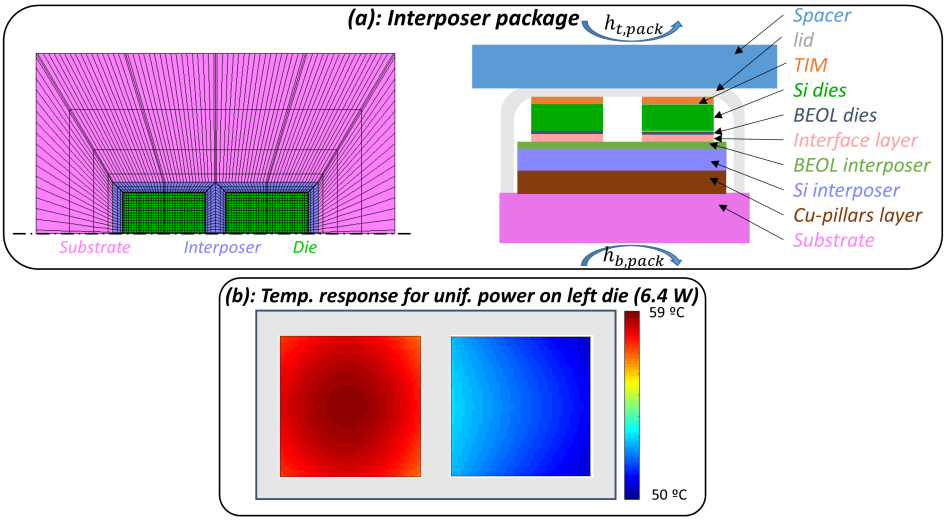


Figure 8.3: (a): Top view and lateral cross section (not to scale) of the geometry used to validate the FTM for interposer configuration. (b): Temperature increase profiles obtained by FEM for uniform power dissipation on the left die. The applied BCs are reported in Section 8.2.2.

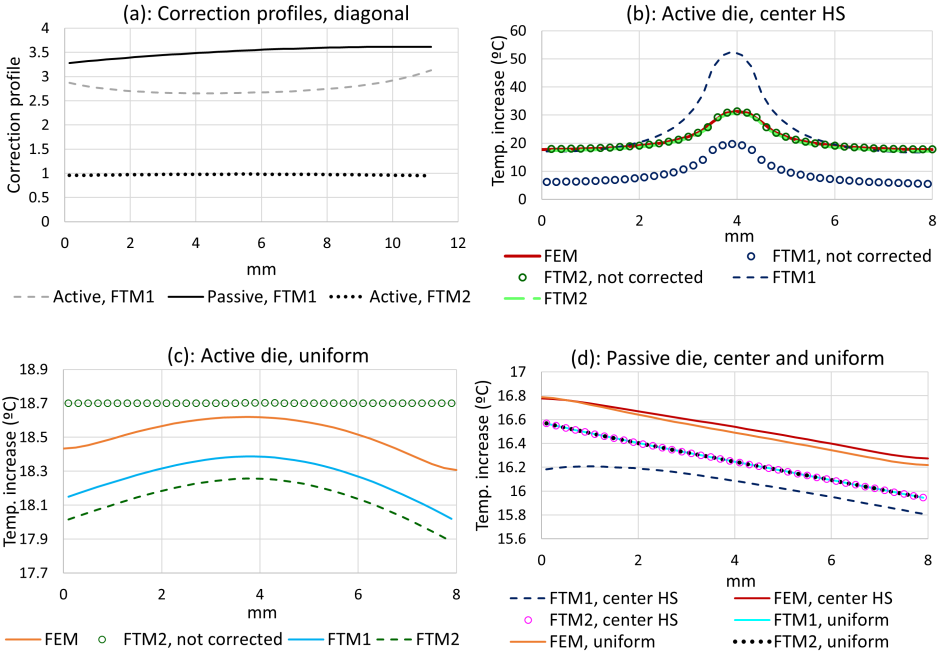


Figure 8.4: Comparison between the correction profiles and the temperature profiles obtained by applying algorithm *FTM1* and *FTM2* for the interposer configuration.

The difference between the two approaches and the impact of the corresponding correction profiles are clearly visible also in the other plots in Figure 8.4. Figure 8.4 (b) reports, in particular, the results on the active die in case of a $1 \times 1 mm^2$ *HS* in the center of the die. The application of algorithm *FTM1* provides a bad approximation, even after having applied the package correction, of the temperature increase and the difference between the results with and without package correction is relevant. For algorithm *FTM2*, instead, the accuracy is good ($\%err < 2\%$ on peak temperature) and the difference between the corrected and the uncorrected results is limited. The same considerations about accuracy are valid for the passive die (Figure 8.4 (d)).

In case of *uniform power dissipation* (Figure 8.4 (c) and (d)), the situation is reversed: after correction, *FTM1* on the active die performs slightly better than *FTM2* (the curve for *FTM1 before correction* is below the values reported in the graph). However, even by applying the *FTM2* methodology, that performs much better in case of non-uniform power dissipation, the error on the maximum temperature remains below 2.3%. Due to the definition of the temperature on the passive die, there is no difference between the results obtained by the two models in this case. This is because the correction profile, in case of *FTM1*, and the temperature increase itself, in case of *FTM2*, are both computed for uniform power dissipation on the other die.

The reason behind the huge difference between the results obtained by the two algorithms and behind the bad approximation obtained by *FTM1* is in how the *thermal spreading within the chip itself* is accounted for in the two models. The silicon is, indeed, a much better thermal conductor than the materials considered around the dies (air, overmold,...). This means that, the assumption of a stack configuration as large as the interposer itself changes the heat path much more than not considering the package at all (cf. Figure 8.4 (a)). The correction profiles are included to take care of the difference in thermal spreading between the interposer and the stack configuration. They are, however, computed for uniform power dissipation. Under this condition, indeed, both modeling methodologies perform good. If HS power dissipation is considered together with algorithm *FTM1*, however, the application of the correction profiles underestimates the spreading *in* the die itself. This is because the correction makes sure that the fictitious spreading in the *not-existing* silicon is removed but, since these profiles are computed for uniform power dissipation, also the real spreading in the silicon, experienced for HS but not for uniform power dissipation, is removed. The opposite is true for *FTM2* and HS power dissipation: the temperature reduction, due to the application of the package correction, is slightly too high (peak temperature increase for the case in Figure 8.4 (b): FEM=31.31°C, *FTM2* w/o correction =31.38°C, *FTM2*=30.71°C). However, since the thermal conductivity of the overmold (or air) is much lower than the one of silicon, the model that starts by considering insulation around the die performs better.

These comments on the correction profiles are further confirmed in Figure 8.5

where the normalized temperature profiles, for the same cases as in Figure 8.4 (b), (c) and (d), are reported. In Figure 8.5 (a), however, all the curves are obtained by means of dedicated FEM models. The correction profiles in Figure 8.5 (b) are obtained as the ratios between the corresponding curves in Figure 8.5 (a). They are, therefore, the *real* corrections needed in each specific situation to meet the temperature increases calculated by FEM for the interposer configuration. In case of the stack geometry considered in *FTM1*, the correction needed for HS power dissipation is much smaller in the center than the one computed by uniform power dissipation, which is the one actually used in the *FTM1*. As a consequence, for high resistive packages, algorithm *FTM1* is not going to work. In case of *FTM2*, instead, the correction needed for HS power dissipation is more or less the same as the one computed for uniform power dissipation. For this case, therefore, this methodology works and it will be considered in the rest of the Section.

8.2.2 FEM validation

In this Section, the validation of the FTM methodology for the interposer configuration is reported. The considered geometry has been described in Paragraph *"Test structure"* in Subsection 8.2.1 (cf. Figure 8.3) while the dimensions and the material properties are listed in Table 8.1. The heat transfer coefficients used for the FEM model are $h_{b,pack} = 95W/mm^2K$ and $h_{t,pack} = 50W/mm^2K$. The corresponding values applied to the stack configuration are $h_b = 2752W/mm^2K$ and $h_t = 1354W/mm^2K$.

Four different power dissipation scenarios are considered. In all cases, one die is active and the other one is kept passive. Three power maps have a HS power dissipation of $1 \times 1mm^2$: external corner, center and internal corner (close to the passive die) of the die. Finally, the results for a uniform power dissipation scenario are also included. In all cases, the total dissipated power is $2W$ (cf. first row in Figure 8.6).

Figure 8.6 reports the results of this validation: on the second row the temperature maps obtained by the FTM are plotted for the four analyzed cases while, on the last two graphs, the results obtained by FEM, FTM and FTM without package correction are compared along the cross section passing through the center of the HS, on the power dissipation level. Orange is used for the results concerning the HS in the outside corner, blue in the center, gray in the inside corner and yellow refers to uniform power dissipation. FEM results are indicated by full lines, FTM results without the correction by the pointed lines and dashed lines are used for the final FTM temperature profiles. As we can see, the FTM approximation is really good in all cases, both for the active and for the passive die. Table 8.2 reports the percentage error in the locations of the maximum temperature on the active die for the considered cases. Due to the small temperature increase on the passive die with respect to the active one, the absolute error is considered in these cases. Also,

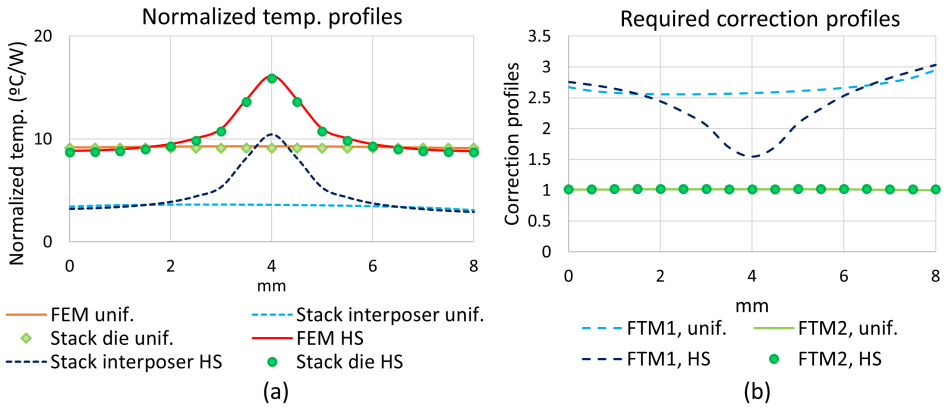


Figure 8.5: (a): Normalized temperature profiles for uniform and HS power dissipation computed by FEM for the real geometry and for the stack configurations considered in *FTM1* and *FTM2* (b): Real correction profiles for the two proposed algorithms in case of uniform and HS power dissipation.

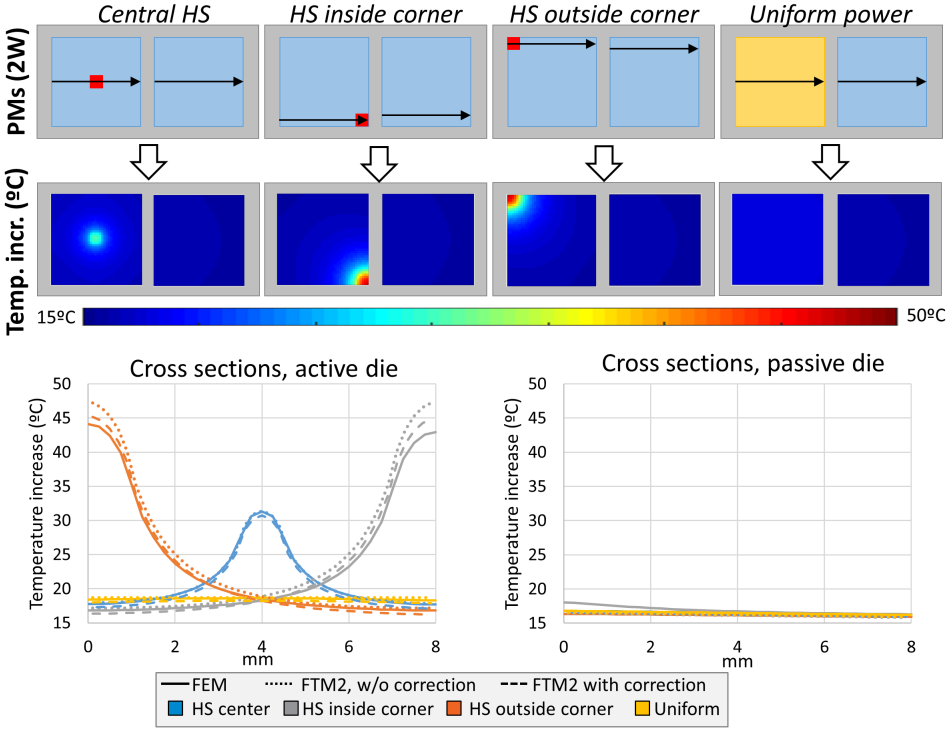


Figure 8.6: Applied power maps and temperature increase profiles obtained by the FTM for the interposer configuration. Comparison of the cross sections of the temperature increase results obtained by FEM, FTM and FTM without package correction.

Table 8.2: Error in the location of the maximum temperature for the cases considered in Figure 8.6.

	External HS	Central HS	Internal HS	Uniform
%*err*, active die, w/o corr.	7.07%	0.21%	10%	0.44%
%*err*, active die, w corr.	2.46%	1.9%	4.42%	1.95%
|*err*|, passive die	0.14°*C*	0.2°*C*	1.6°*C*	0.22°*C*

due to the definition of the FTM for the passive die, just the "corrected" results are available.

The numbers in Table 8.2 show that the accuracy of this extension is good. The interposer package correction is particularly efficient if the HS are dissipated in the corners and, as explained in the previous Subsection, slightly worsens the results for HS power dissipation in the center. The same happens for uniform power dissipation but, due to the small temperature increase in this case, the worsening, presented as percentage errors in Table 8.2, corresponds to an absolute error variation from 0.22°*C*, without correction, to 0.36°*C*, with correction, which is definitively acceptable.

In all these analyzed cases, one die is always considered passive. In case of situations in which both dies are active, the application of superposition will provide the required temperature estimation. The accuracy of the temperature profile on the passive die is, in these cases, even less relevant because of the higher impact of the self-heating of each die.

The results shown in this Section prove that the algorithm, presented as *FTM2*, is able to accurately predict the temperature increase in case of an interposer geometry for both uniform and non-uniform power dissipation.

8.3 Stack of dies with different sizes, transient regime

In this Section the FTM is extended to a *pyramidal configuration*, i.e. a structure constituted by a stack of dies with different sizes. This situation can occur, for example, in case of heterogeneous integration. This kind of 3D-ICs are build by stacking dies that perform different functions (logic and memory, for instance) and/or that are manufactured in different companies, meaning that they might have different specifications and dimensions.

Transient experimental results were obtained at imec for a specific test chip (named *3D130c*) with these characteristics [76, 103]. This device mimics hot spot power dissipation of a real chip and, at the same time, monitors the transient temperature increase in specific locations. In order to be able to experimentally validate the

model, the study of the applicability of the FTM methodology to a pyramidal geometry has been performed directly considering this specific test chip. The methodology is, however, not restricted to this specific structure and it can be extended to other pyramidal configurations. What has not been developed, is the extension of the model to situations in which, for a pyramidal geometry, power is dissipated on the largest die.

The validation of this extended FTM is initially performed with respect to FEM. This allows to check more easily if the algorithm itself is applicable, without any disturbance from measurement errors and variability, as well as from structures which are not included in the FTM (TSVs, μbumps, BEOL, ...). For this initial validation, a simplified version of the 3D130c test chip is considered. The next Subsection is dedicated to the description of the test vehicle, afterwards the FEM validation is presented and, finally, the experimental validation is described.

8.3.1 Test chip: 3D130c

The 3D130c test chip consists of two dies stacked on top of each other. The $5\times5mm^2$ top die, which is thinned down to $25\mu m$, is stacked faced-up (F2B configuration) on top of a $8 \times 8mm^2$ full thickness ($725\mu m$) bottom die (cf. Figure 8.7). Multiple heating and temperature sensors configurations are designed in this test vehicle. However, just the one relevant for this work is described here. For an in-depth description of the test vehicle, please refer to [76, 102, 103].

The power dissipation of a real device is mimicked by a $100 \times 100\mu m^2$ Cu heater in the metal 2 layer of the BEOL of the top die. A significant temperature drop is expected to occur between the heater and the top silicon die. This is because the metal layers of the BEOL are separated from the top silicon die by a SiO_2 layer with low thermal conductivity. For this reason, the temperature in the heater is supposed to be significantly higher than the one measured by the temperature sensors, which are located in the active regions of the top and bottom silicon dies, and reported in this Section. Diodes are placed in correspondence of and at various distances from the heater's center. Figure 8.7 (a) shows the floorplan of the test vehicle. The location of the heater is highlighted by a red circle.

8.3.2 FEM validation

As mentioned before, to check the applicability of the developed transient FTM, the obtained results are initially compared with the ones from standard FEM techniques. To achieve this aim, the geometry of the 3D130c test vehicle has been simplified in the FEM so that exactly the same structure can be analyzed by the FTM and the FEM. This means that all the small structures, as TSVs, μbumps, metal lines,..., are not modeled individually neither in the FEM nor in the FTM

Table 8.3: Parameters used in the FEM and FTM simulations of the stack of dies with different sizes.

Layer (from top)	Thickness (μm)	$k(W/mK)$	$c(J/kgK)$	$\rho(kg/m^3)$
SiO_2	0.83	1.4	1000	2200
	Power dissipation			
Cu	0.7	401	385	8960
SiO_2	1.3	1.4	1000	2200
	Temperature response			
Si	21.05	148	700	2330
Interface	0.5	0.29	1178	1060
SiO_2	2.05	1.4	1000	2200
	Temperature response			
Si	725	148	700	2330

geometry. Moreover, since one of the main assumption of the FTM is that the HSRs are position independent, a full layer of Cu is assumed on top of the top die to model the metal layer of the BEOL. In the real device, Cu is localized where the power is dissipated but this creates a situation where different materials are present on the same horizontal layer, which cannot be directly handled by the FTM. The schematic of the vertical cross section of the modeled device is shown in Figure 8.7 (b). In Table 8.3 the values of the different parameters are listed. A mesh size of $20\mu m$ is used and insulation is assumed everywhere, except for the bottom of the stack where a convective boundary condition is applied. This is because, in the real experimental setup, the test vehicle is placed on a temperature controlled chuck (15°C) and, as a consequence, the heat is mainly removed from the bottom of the stack. However, since the chuck is not included in the FEM, the value of h, which represents an equivalent heat transfer coefficient, is selected so that it accounts for the thermal effect of both the chuck and the contact resistance between the test vehicle and the chuck itself, resulting in $h_{b,pack} = 1500W/m^2K$.

The next step is to adapt the FTM developed for stacks of dies with the same footprint area to this new situation. The idea is to consider the larger bottom die in a similar way as the package thermal spreading is handled in a conventional 3D-package. This means that all the material layers with a surface area of $5 \times 5mm^2$ (first four layers starting from top) are assumed to play the same role as what was previously considered the *die stack*, while the larger lower layers are considered to act as the *package*.

In Section 5.4, where the transient package correction was introduced, a *fixed time step* was considered. In Section 7.4.3, the possibility to use a *variable time step* has been presented in case information over the temperature evolution at short time is needed. In both cases, the data concerning both the HSRs and the correction profiles are stored at the selected points in time. The use of a constant, or of a

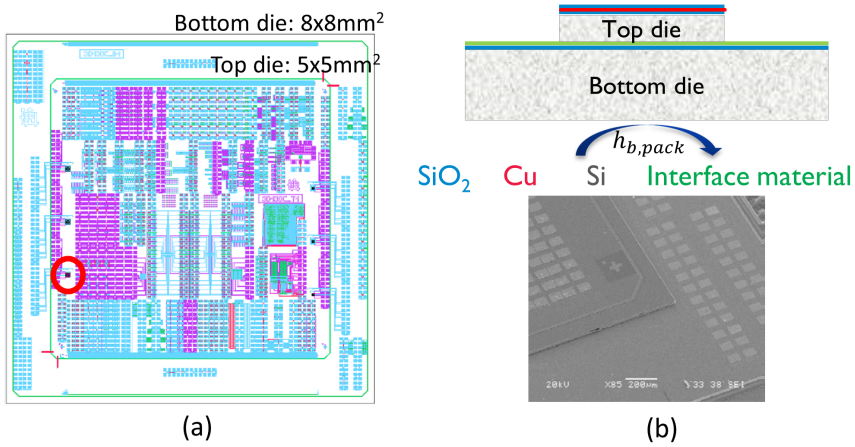


Figure 8.7: (a): Floorplan of the 3D130c test vehicle. The results shown in this Section refer to the location of the $100 \times 100 \mu m^2$ heater indicated by the red circle. (b): Schematic of the vertical cross section of the geometry used in the FEM and FTM models (not to scale, $h_{b,pack} = 1500 W/m^2K$) and picture of the fabricated device [76].

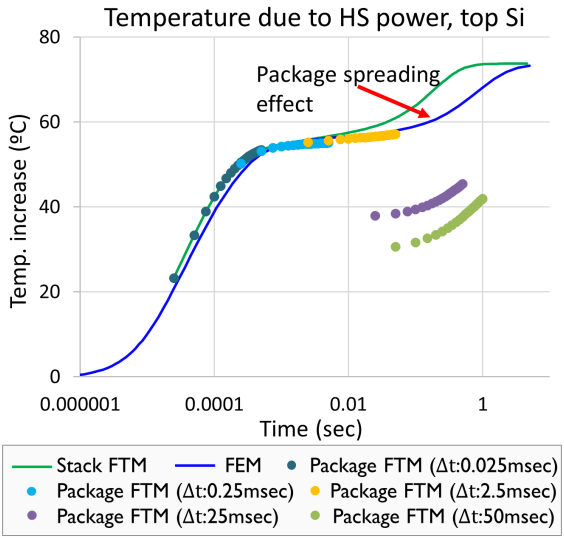


Figure 8.8: Transient temperature response in the center of the HS on top of the top die for the pyramid configuration in Figure 8.7. The green line is for the stack FTM and the blue line for the FEM results. Dots represent the results obtained by the package FTM with different constant time steps.

large, time step in the model for the stack geometry is possible because these data are originally computed by the FEM software using an adaptive time step, which allows to correctly follow the fast variation of the temperature during the heating up and cooling down phases. Since these fast variations are already considered in this FEM phase, they don't need to be followed again for each specific power map and the selection of the time steps can be independent of the system itself. This is perfectly true if the FTM without package correction is considered (green curve in Figure 8.8). In this case, whatever time step is selected, the obtained results agrees with the FEM ones (blue curve in Figure 8.8) until $10msec$. The difference between these two curves starting at $10msec$ comes from the spreading due to the larger size of the bottom die, which is not included in the transient stack FTM. However, as can be seen from the dots in different colors in the same plot, the results for the package FTM can differ a lot from the related FEM if the selected time step is too large. In all the reported cases, constant time steps are used and, if $\Delta t \geq 25msec$, the error becomes large. As long as the time step is smaller than the time where the spreading and the capacitive effects become significant, the package FTM works fine and it is able to include the thermal spreading due to the larger bottom die. This means that, if a small enough time step is selected, the package FTM methodology can be used to properly model this structure.

The reason behind this issue concerning the time step is in the difference between the heating curves for the stack and for the package configurations in case of uniform power dissipation. Their ratios at each time step are, indeed, used as correction profiles. Figure 8.9 shows the time evolution of the heating curves in the location of the maximum temperature for a pyramidal structure (lines without markers, left axis) and for a conventional 3D-package (lines with markers, right axis). Since the thermal diffusivity ($k/\rho c$) of Si is much higher than the one of conventional packaging materials (overmold, substrate,...), the difference between the heating curves for the stack and the package configuration starts earlier in time in case of the pyramidal geometry. The issue in the FTM arises because, being the time in the FTM discrete, the correction profile computed at a specific time step is applied to the *whole* previous time interval. This means that if, for example, a time step of $50msec$ is considered, the same difference between the heating curves experienced at $50msec$ is assumed from the beginning of the heating process until $50msec$. As a consequence, the corresponding correction profile is assumed to be valid and constant for the whole time step. Which is not the case. It is important, therefore, to consider a time step that is smaller than the point in time where the stack and the package temperature responses start to deviate significantly. Since these two heating curves are available from, respectively, the application of the *annulus method* to the stack configuration (cf. Paragraph *"Temperature for uniform power dissipation in the stack configuration"* in Section 5.4.3) and from the FEM results of the coarse package simulation, this information is readily available for their comparison.

Having a constraint on the time step in the transient package FTM could affect computational time and cause memory issues if the time, needed by the system to

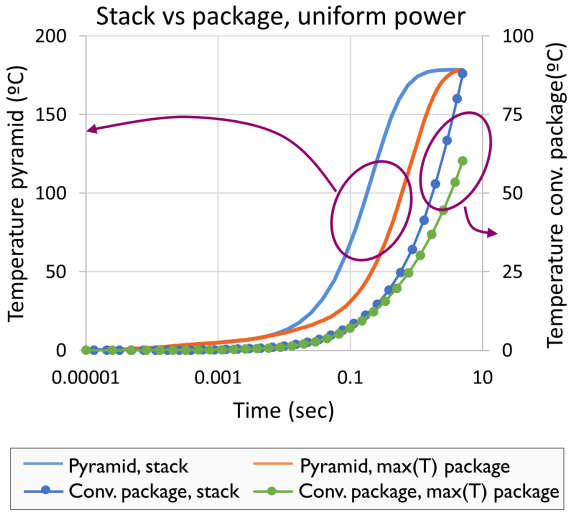


Figure 8.9: Transient temperature responses in the location of $max(T)$ for uniform power dissipation. Comparison of the response of the package and the stack configurations in case of a conventional 3D-package (right axis, lines with markers) and of a pyramidal configuration (left axis, line without markers).

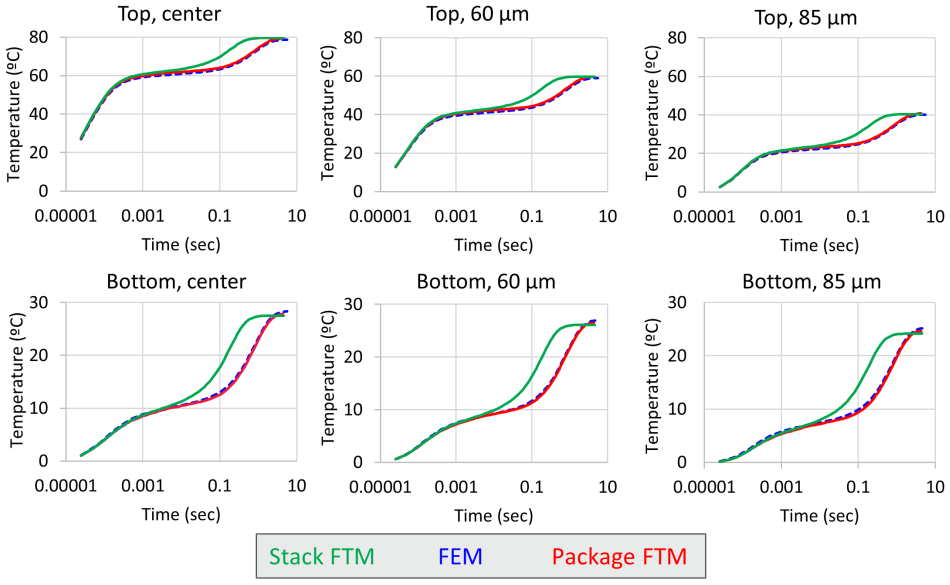


Figure 8.10: Transient temperature responses to hot spot power dissipation on top of the top die (first row) and of the bottom die (second row) for the pyramid configuration in Figure 8.7. Results are reported in correspondence of the center of the HS, at $60\mu m$ distance along the horizontal direction and at $60\sqrt{2}\mu m$ distance along the diagonal direction.

reach steady state, is much larger than the required time step and if a constant time step is used. For this reason, the procedure based on the variable time step has been applied, even if we are not interested in short time information. The reason underlying the validity of a variable time step approach in this situation is the same as the one described in Paragraph *"Temperature profiles extraction"* in Section 5.4.3, which justified the choice of a scaling technique, instead of the extraction of FEM results at each time step, to compute the correction profiles. The longer ago an impulse has been dissipated, indeed, the lower its impact on the actual temperature. The corrections can, therefore, be less accurate at later times.

The validation of this algorithm with respect to FEM is shown in Figure 8.10. A $100 \times 100 \mu m^2$ HS power dissipation is assumed on top of the Cu layer and the temperature responses are reported on top of the top (first row) and of the bottom (second row) silicon die. Results are reported in correspondence of the center of the HS, at $60 \mu m$ distance along the horizontal direction and at $60\sqrt{2} \mu m$ distance along the diagonal direction in the first, second and third column, respectively. Green is used for the stack FTM, blue for FEM and red for the package FTM with variable time steps. The initial time step is set to $25 \mu sec$. Every 20 steps, the time step is multiplied by ten, until the value of $4.5555 sec$, which is the point in time when the simulation ends, is reached. The agreement between the package FTM and the FEM results is really good from the beginning to the end of the process, fully validating the applicability of this package approach to the pyramidal geometry once a proper time step has been selected.

8.3.3 Experimental validation

Experimental transient results are available on top of the top and of the bottom die aligned with the center of the HS. The thermal measurements are performed on a wafer level probe station. The die stack is mounted on a temperature controlled chuck at 15°C. Probes are put in contact with the bond pads of the top die to provide power to the heater, to force current in the top and bottom diodes and to measure the correspondent voltage. The distance between bond pads and diodes is sufficiently large ($\approx 1 mm$) to avoid any impact of the former on the thermal behavior of the structures. The transient temperature response of the system is monitored by the embedded sensors allowing to characterize the temperature evolution in time [76].

The most important assumption underlying the stack FTM is that all the vertical cross sections of the stack configuration are the same. Fortunately, lot of exceptions to this rule can be neglected without having a significant impact on the final results (metal lines, TSVs, μbumps...). This is especially true if these exceptions are far away from the power dissipation position. Unfortunately, in this test vehicle, the heater, which can be simplified as a Cu block, is surrounded by SiO_2, which is a bad thermal conductor. Copper is, therefore, limited to the heater area, it is

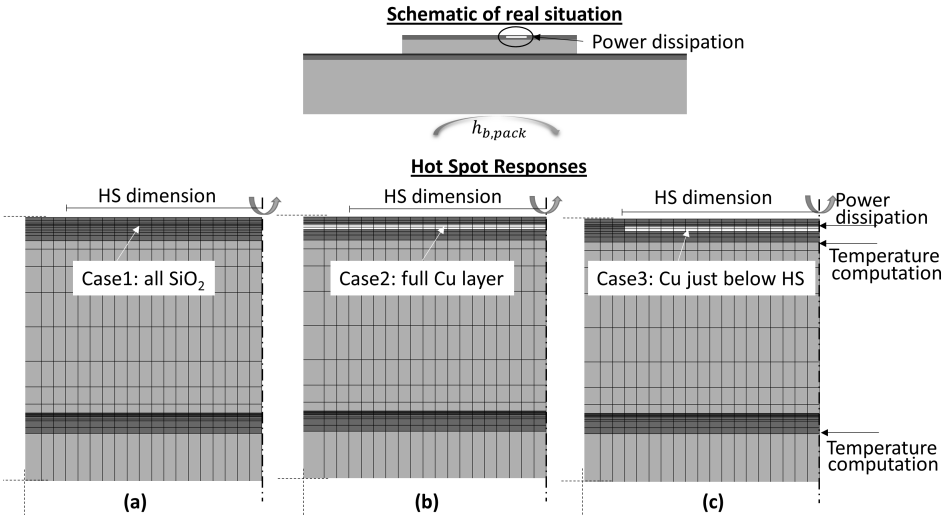


Figure 8.11: First row: schematic of the real configuration with the Cu block located aligned with the power dissipation position. Second row: three possible geometries for the HSRs; (a) no Cu considered (just SiO_2), (b) full layer of Cu and (c) Cu just aligned with the HS position.

not a full layer. Due to the proximity of this material heterogeneity to the power dissipation location and due to the difference in thermal diffusivity between SiO_2 and Cu, the impact on the temperature profiles is significant.

In order to stick to the basic assumptions, two options can be considered: 1) to neglect the Cu and to consider a full layer of SiO_2 or 2) to consider a full layer of Cu, as in Section 8.3.2. The geometries of the corresponding HSRs are reported in Figure 8.11 (a) and (b), respectively. Due to the small area of Cu with respect to SiO_2 material, the option of using equivalent material properties is comparable to the one of a full layer of SiO_2. The results obtained considering these two geometries are shown, respectively, in gray and orange in Figure 8.12. The light blue dots are the experimental results. The graph on the left refers to the top die while the one on the right to the bottom die. As foreseeable, the impact of the selected material is quite relevant, especially in the top die, due to its vicinity to the heat dissipation location. Furthermore, the experimental results, which are obtained for a situation that can be simplified as somewhere in between the two modeled ones, lies in between the two curves. The lower impact of the chosen material on the temperature of the bottom die is due to the longer heat path from the heat dissipation position to the temperature response level. Since other materials enter the heat path, the relevance of this small heterogeneity becomes less significant. Moreover, in this farther away location, other structures, which have been neglected in the FTM, play also a more significant role in the error between experiments and FTM.

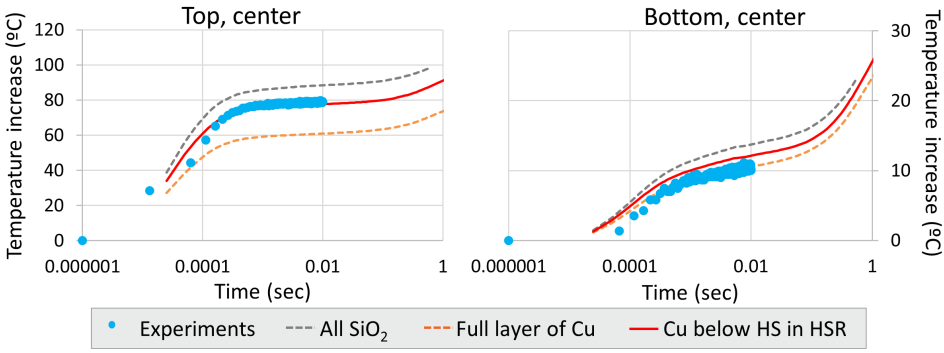


Figure 8.12: Temperature responses for HS power dissipation on top of the top die (left) and of the bottom die (right) in the center of the HS. Light blue dots refer to experimental results, gray curves to the packaged FTM assuming no Cu at all, and orange curves to the packaged FTM assuming a full layer of Cu. The red curves are obtained by considering a small area of Cu just below the HS while computing the HSRs.

A third approach that can be considered consists in *violating*, on purpose, the basic assumption of homogeneous material horizontal layers. Since the Cu material is placed just below the power dissipation area, we could try to calculate the HSRs assuming Cu just below the HS and SiO_2 elsewhere (cf. Figure 8.11 (c)). Due to the specificity of this situation, where Cu and power are localized and coupled, this algorithm produces good agreement with the experiments (red curve in Figure 8.12). Concerning the correction profiles, since they are obtained for uniform power dissipation, a full layer of Cu is considered. With this trick, valid for this specific test vehicle or for similar situations in which a better thermal conductor is coupled with power dissipation, the packaged FTM has been validated with respect to experimental results.

8.4 Summary

In this Chapter, the FTM previously presented for 3D-packages has been successfully extended to two other different configurations: *interposer* and *pyramidal stacks*.

Two possible FTM approaches have been assessed for the thermal analysis of the interposer configuration in the steady state regime. One of the two revealed to be unsuitable to model the considered situation in case of non-uniform power dissipation. The other one, instead, proved to be accurate for both HS and uniform power dissipation and it has been successfully validated with respect to FEM results in case of a packaged interposer. The temperature profiles are reported on the

active layers of the active dies but, if the HSRs are extracted also in correspondence of the interposer, the temperature profiles can be computed also in this location. Moreover, the presented FTM assumes a passive, silicon interposer die. It can be expected that the same methodology also works in case of a glass or of a laminate passive interposer die. For an active interposer, instead, a constriction resistance (from the larger interposer die to the smaller dies on top of it) should probably be considered. A similar methodology as the one presented in this Chapter could probably be applied, but further research is needed.

The FTM presented for the pyramidal geometry (stack of dies with different sizes) has not only been successfully validated with respect to FEM results, but also with respect to experiments. Since experimental results were available in the transient regime, the extended FTM has been developed in this regime. Results in steady state are not reported because they represent a special case of the transient regime. Also for this configuration, in case power is dissipated on the larger die outside the area corresponding to the smaller top die, further research is needed to check the impact of the thermal constriction.

A common point between these two extensions is that the difference, with respect to the original 3D-package FTM, is in the computed package correction. In the interposer configuration, for example, everything outside the die stack of a single die is considered as package, resulting in a non-symmetric package correction including also the region of the second die. For the pyramidal geometry, instead, part of the larger silicon die is considered as package. Due to the much higher thermal diffusivity of silicon with respect to the one of the materials used for conventional 3D-packages (substrate, overmold,...), care should be taken in the selection of a small enough initial time step. The success of these extensions proves that the concept of the *package correction* is not restricted to 3D-package configurations but it can also be extended to different situations.

Chapter 9

Applications & Case Studies

9.1 Introduction

In this Chapter, several potential applications of the FTM are presented. These case studies are related to realistic situations that may occur during the design phase of an IC. The use of the FTM illustrates how the thermal analysis can be performed in an efficient way for a wide range of situations. In particular, since for newly developed stacking technologies thinned dies are used, the applicability of the FTM to the analysis of the thermal impact of die thinning is presented in Section 9.2. The thermal performance of a realistic power map (OpenSparc) will be analyzed considering dynamic power dissipation in case of a 3D configuration (Section 9.3.1) and by comparing the thermal performances of a 2D and a 3D technology option (Section 9.3.2). Finally, the thermal impact of different interface materials is analyzed in Section 9.4.1 and the one of different layouts and amount of μbump arrays in Section 9.4.2.

9.2 Thermal impact of die thinning

Extremely thinned dies are used in 3D-SOC (system-on-chip) integration [9,28]. Following the evolution of the technology, dies are thinned down to $5\mu m$ (or even less) and the consequences of this thinning have to be considered and analyzed also from a thermal point of view. In a simple case, with *uniform* heating and one directional heat flow through a single layer of material, the thermal impact of die-thinning can be computed straightforward. The temperature increase can, indeed, be calculated as $\Delta T = \frac{l}{kA}Q$. This means that, by reducing the thickness l of the die, the temperature increase is also reduced and a thermal improvement is,

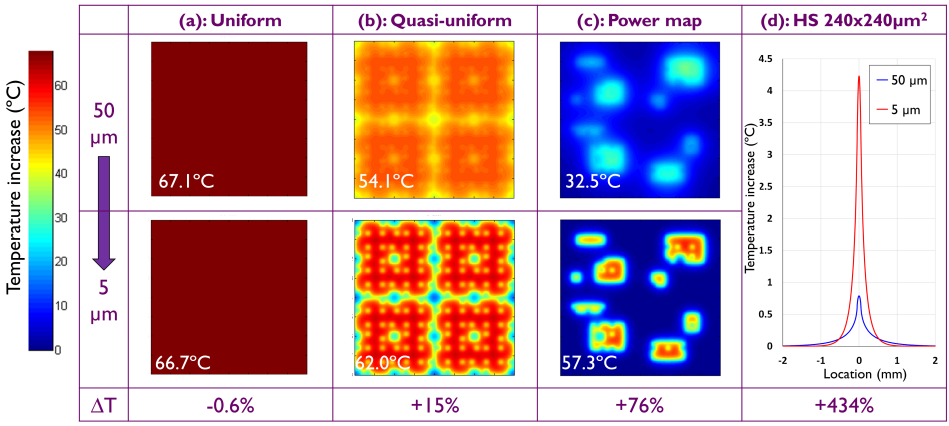


Figure 9.1: Temperature maps in case of die thinning for different power dissipation scenarios.

therefore, achieved. However, due to the small impact of the silicon layer to the overall thermal resistance, the temperature reduction achieved by die thinning is very limited. Moreover, a real scenario rarely involves uniform power dissipation. The impact of die thinning drastically changes for *non-uniform* power dissipation (in case of convective BCs). In this latter case, indeed, heat spreading plays an important role: thinner dies mean less room for lateral heat spreading and higher temperature peaks. The worsening of the thermal behavior depends, in particular, on the size of the HS, on the thermal resistance of the material and on the applied boundary conditions.

The FTM has been applied to evaluate the thermal impact of die thinning for several cases. A simple model consisting in just one silicon layer has been considered. The temperature profiles obtained in case of a $50\mu m$ thick die are compared with the ones obtained for a $5\mu m$ thick die, for different power dissipation scenarios. In all cases, the power density is fixed to $1W/mm^2$ and the heat is convectively removed from the bottom side of the structure ($h_b = 15000W/m^2K$).

Figure 9.1 shows the different temperature responses for the different power dissipation scenarios. Results concerning the thick die are reported on the top row while the ones concerning the thin die on the bottom row. The maximum achieved temperature in each situation is reported on the plot itself. Let's first consider the two most extreme cases: uniform power and HS power dissipation. For uniform power dissipation (left hand side), as already mentioned in the previous paragraph, a small drop in temperature increase from 67.1°C to 66.7°C (-0.6%) is experienced by thinning the die. For HS power dissipation (right hand side), instead, the peak temperature strongly increases from 0.8°C to 4.3°C (+434%) by performing the same operation. From this graph, a second effect of the die thinning is visible: the higher curve is also narrower. This is the reason why a

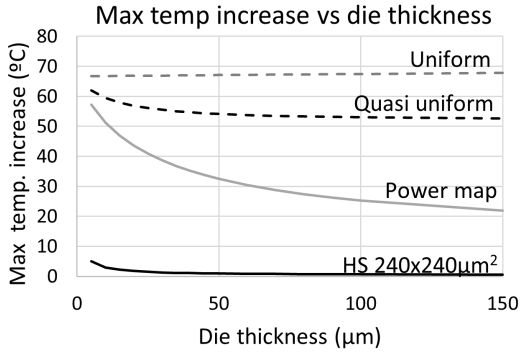


Figure 9.2: Maximum temperature increase as a function of the die thickness for the different power dissipation scenarios shown in Figure 9.1.

HSR with a higher peak generates a lower temperature increase in case of uniform power dissipation. By performing superposition (or convolution), in case of a thinner die, less overlapping is experienced between neighboring curves and, as a consequence, the overall temperature increase is lower.

The other two graphs in Figure 9.1 show the results in case of other non-uniform power dissipation scenarios. The considered PMs are related to the PTCQ test vehicle presented in Chapter 7 for the experimental validation of the FTM. The temperature profiles in column (b), in particular, are obtained assuming all heaters on (cf. Figure 7.1) while, for the temperature profiles in column (c), the power map in Figure 7.6 is applied. A general comment is that the thinner the die, the *more defined* the high temperature areas are; the thicker the die, the more *blurred* temperature maps are obtained. Furthermore, a higher difference between the maximum and the minimum achieved temperatures is experienced in case of thin dies and non uniform power dissipation. All these effects are due to the smaller correlation between neighboring cells.

Figure 9.2 reports the maximum temperature increase as a function of the die thickness (from $5\mu m$ to $150\mu m$) for the different power dissipation scenarios considered in Figure 9.1. The graph clearly shows that the dependence of the maximum temperature on the thickness of the die is highly dependent on the dissipated power map. In case of the specific power map (case (c) in Figure 9.1), for instance, the impact of die thinning is much more evident than in case of quasi uniform power dissipation (case (b) in Figure 9.1).

This kind of analysis can be easily performed by the developed FTM. The methodology is, indeed, able to provide quick and accurate estimations of the temperature increase given specific thicknesses of the dies and PMs. This is possible by simply convolving appropriate HSRs with the desired power maps. In this way, multiple scenarios can be easily compared, helping in selecting an

Table 9.1: Values of the parameters used in the FTM simulations in Subsection 9.3.1.

Parameter	Value	Parameter	Value
cs	$5 \times 5mm^2$	c Si	$700J/kgK$
l_t	$250\mu m$	c interface	$2187J/kgK$
l_{int}	$13\mu m$	ρ Si	$2330kg/m^3$
l_b	$50\mu m$	ρ interface	$1051kg/m^3$
k Si	$120W/mK$	k interface	$1W/mK$
R_{th} top-to-ambient	$0.8K/W$	R_{th} bottom-to-ambient	$2000K/W$
l "package"	$1mm$	$c \cdot \rho$ "package"	$11000 \cdot 1051J/m^3K$
Δt	$50msec$	$\bar{h}$	$100\mu m$
t_f	$3sec$	T_{amb}	$25°C$

appropriate solution.

9.3 Applications for the OpenSPARC floorplan

9.3.1 Dynamic power dissipation

The second test case concerns the transient analysis of a *memory on logic* die stack, for which a 3D repartitioning of the open source OpenSparc T2 floor plan [79,97] into two smaller stacked dies is considered (Figure 9.3). The results presented in this Subsection have been published in [62]. The modeled geometry is similar to the one sketched in Figure 3.12. With respect to that sketch, a layer of material is added on top and bottom of the stack in order to include the capacitive effect of the package (without including its spreading impact). The heat is mainly removed from the top side of the configuration and the parameters used in the simulation are listed in Table 9.1. The logic die is assumed to be the bottom one in the stack, while the memory die the top one. This case study is similar to the example that considers duty cycles in the experimental validation of the FTM (cf. Section 7.4.3). In this case, however, the power map is related to a more realistic scenario and the coupling between the two dies is also included.

Three different scenarios have been analyzed:

Scenario1: Both the memory and the logic die are 100% on for the whole simulated time;

Scenario2: The memory is always on while, for the logic die, the load is switched from the four central cores (2, 3, 4 and 5) to the four external ones (0, 1, 6

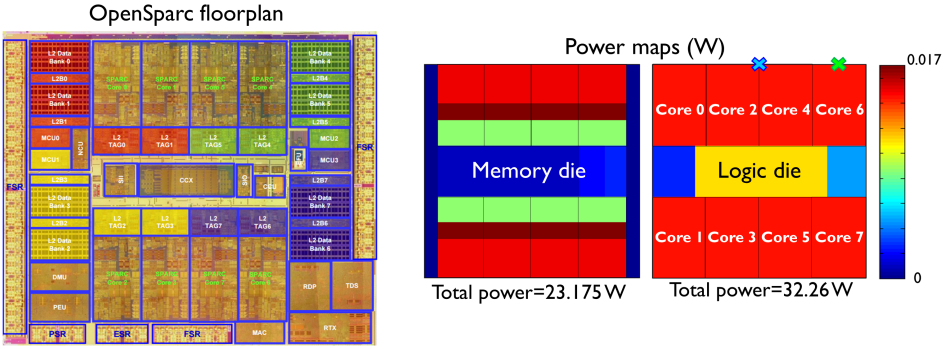


Figure 9.3: OpenSparc floor-plan [98] and power distribution on the memory (top) and the logic (bottom) die.

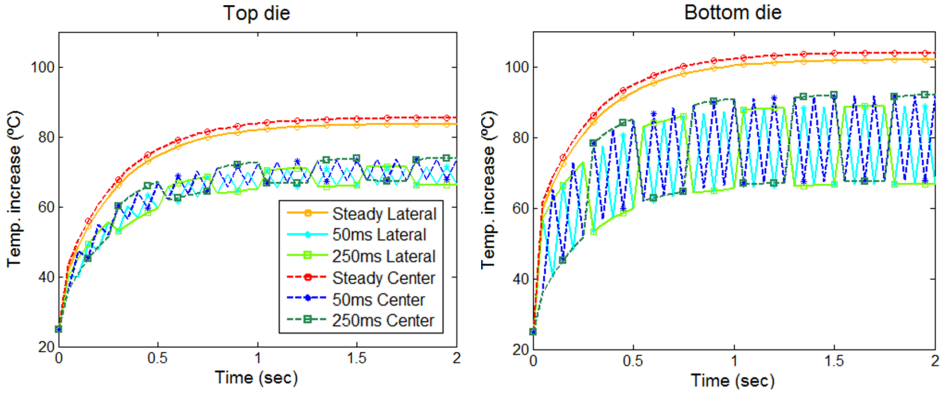


Figure 9.4: Temperature evolution as a function of time in the points indicated with a light blue (label *center*) and green (label *lateral*) cross in Figure 9.3. Top (memory) die values on the left and bottom (logic) die values on the right.

and 7) every 50 msec. When a core is on, it dissipates 100% of the power (as represented in Figure 9.3), while, when it is off, it dissipates 0W;

Scenario3: This scenario is similar to Scenario2 but the switching of the loads between the cores happens every 250 msec instead of every 50 msec.

Figure 9.4 shows the time evolution of the temperature values in the places indicated, in Figure 9.3, with a light blue cross (labeled by *center* and represented by darker tones in Figure 9.4) and with a green cross (labeled by *lateral* and represented by lighter tones in Figure 9.4). The temperature values referring to the central location are a bit higher than the lateral ones because that point is located in the middle of two active cores while the latter one is in the middle of one active core (the other one is located on the opposite corner).

The curves referring to Scenario1 (circular markers) show an increasing temperature value until steady state is reached. This is, of course, because the power maps are static. The other curves, instead, show a periodic behavior, the only difference being in the period: for Scenario2 (star-shape markers) it is 50 msec while, for Scenario3 (square markers) it is 250 msec.

This is a further prove that the FTM is able to predict the thermal effect of the temporal changes in power dissipation. It can, therefore, be exploited in similar cases when the load can be switched between different cores and an appropriate switching time needs to be calculated.

9.3.2 2D vs 3D technology

A third application concerns the comparison between a 2D and a 3D technology option in steady state regime. The realistic OpenSparc floor-plan is still considered and the parameters used in the two cases are listed in Table 9.2. It is worthy to note that the same functional blocks are present in both cases: in the 3D configuration, they are placed on two levels while, in the 2D configuration, on the same plane.

A couple of words should be spent on the values of the heat transfer coefficients. They have been selected in such a way that the same cooling solution is assumed to be applied on top of both the 2D and the 3D case. If, in particular, a HP configuration with heat sink is considered, the same heat sink is assumed to be placed on top of both the 2D and 3D option. The process to select adequate heat transfer coefficients consists in several steps that are listed hereafter and illustrated in Figure 9.5.

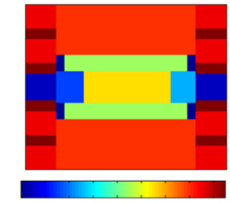
1. Starting from the value of the heat transfer coefficient applied on top of the 3D stack, $h_{t,3D}$, the thermal resistance from top die to ambient, defined as $R_{th,t} = 1/(h_{t,3D}A)$, where A is the floor-plan area, is computed. In particular, from the modeled stack structure, it results $R_{th,t,3D} = 0.93K/W$.

2. A FEM model with a large block of copper ($5 \times 5 \times 1cm^3$) on top of the die stack has been built and run. This large block mimics a generic cooling solution. Uniform power, for a total value of $1W$ is dissipated on top of the $5 \times 5mm^2$ stack.

3. The value of the uniform heat transfer coefficient, $h'_{t,3D}$, applied on the top of this model is selected in such a way that the temperature increment in the stack is $\Delta T_{3D} = 0.93°C$. This is because, being the total dissipated power $1W$, the temperature increase equals the thermal resistance.

4. A convective BC with $h'_{t,3D}$ is then applied on top of the corresponding FEM model for the 2D configuration. This model consists in a $6.2 \times 7.2mm^2$ silicon die with a $5 \times 5 \times 1cm^3$ copper block on top, mimicking the cooling solution.

Table 9.2: Parameters used for the comparison of the thermal performances of the 2D and the 3D configuration in case of the OpenSparc floor plan.

Parameter	2D configuration	3D configuration
cs	$6.2 \times 7.2mm^2$	$5mm^2$
l_t	$250\mu m$ ($13\mu m$ of Si on top)	$250\mu m$
l_b	-	$50\mu m$
l_{int}	-	$13\mu m$
$\bar{h}$	$100\mu m$	$100\mu m$
k Si	$120W/mK$	$120W/mK$
k interface	-	$1W/mK$
T_{amb}	$25°C$	$25°C$
h_t	$27042W/m^2K$	$42857W/m^2K$
h_b	$11W/m^2K$	$20W/m^2K$
PM		

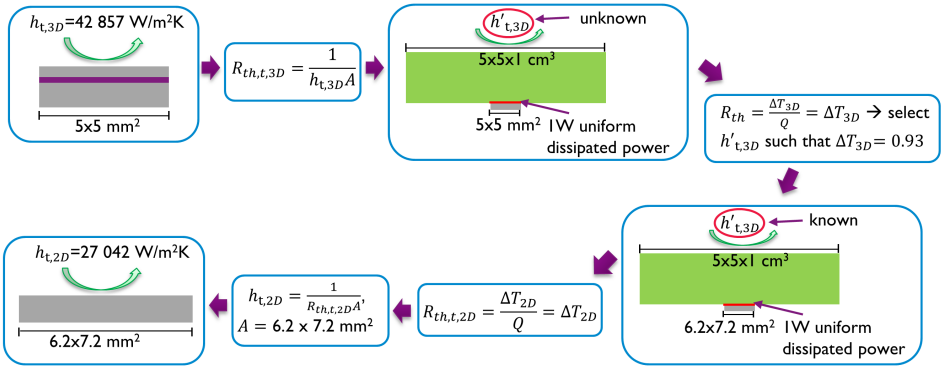


Figure 9.5: Procedure to compute the heat transfer coefficient on the top of the 2D configuration in such a way that the same cooling solution is assumed for the 3D and the 2D case.

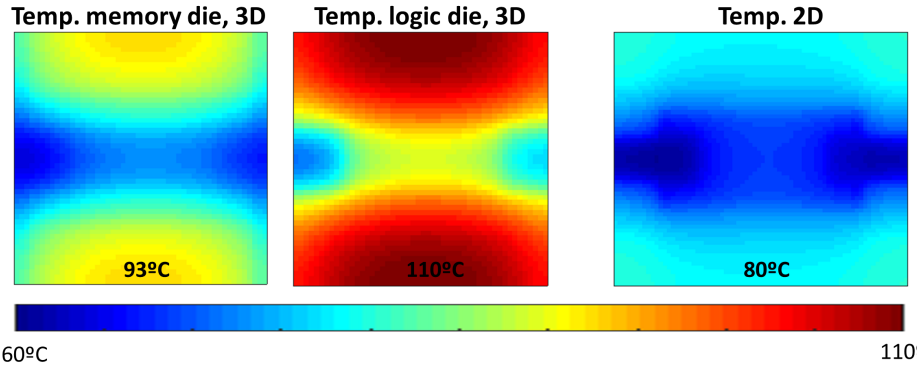


Figure 9.6: Temperature profiles on the memory (top) and on the logic (bottom) die in the 3D configuration and on the 2D configuration. The maximum temperature in each die is also reported.

5. The corresponding $R_{th,t,2D} = \Delta T_{2D}/Q$ is computed from the obtained temperature value.

6. The heat transfer coefficient to be applied on top of the 2D stack configuration is, finally, computed as $h_{t,2D} = 1/(R_{th,t,2D}A)$.

In this way, therefore, the effect of the difference in heat spreading, from the die to the heat sink, is accounted for. Performing just an area scaling, meaning that the same thermal resistance is assumed from the top surface of the stack to the ambient, would have resulted in $h_{t,2D} = 24000W/m^2K$. Concerning the heat transfer coefficients applied on the bottom of the stack, since the heat is mainly removed from the top, a simple scaling with respect to the floor-plan area has been applied (same resistance assumed from the bottom of the stack to ambient).

The temperature maps obtained for these two cases are shown in Figure 9.6. Normally stricter thermal constraints are required for the memory die than for the logic die. For the analyzed case, the maximum temperature in the memory die in the 3D stack is around 93°C, while, in the 2D configuration, in correspondence with the memory functional blocks it is around 80°C (20% difference on the temperature increase). This means that, if the constraint on the maximum temperature for the memory die is around 80°C, in order to have a working device, the cooling solution applied on top of the 3D stack should be improved, from $R_{th,t,3D} = 0.93K/W$ to $R_{th,t,3D} = 0.75K/W$.

This kind of analysis can be easily implemented by the developed FTM. The methodology can, therefore, help in quickly comparing different technology options and to propose adequate cooling solutions in order to avoid temperature driven chip failures.

Table 9.3: Parameters values used in the study of the thermal impact of the properties of the interface material.

	Case1	*Case2*	*Case3*
$k_{\mu b,xy}$ (W/mK)	1.3	1.6	2
$k_{\mu b,z}$ (W/mK)	5	5	7
k_{und} (W/mK)	0.7	1	1.5
k_{Si} (W/mK)	120		
h_t (W/m^2K)	3000		
h_b (W/m^2K)	Insulation		
l_t (μm)	200		
l_b (μm)	50		
l_{int} (μm)	13		
T_{amb} (°C)	25		
$\bar{h}$ (μm)	120		
cs (mm^2)	8.16×8.16		

9.4 Thermal impact of die-die interface

9.4.1 Thermal impact of die-die interface material

A fourth illustration of the capability of the FTM concerns the analysis of the thermal impact of the interface materials. The results presented in this Subsection have been published in [58]. Three different cases, *Case1*, *Case2* and *Case3*, in which different underfill materials and different μbump array equivalent properties, obtained by modifying individual μbump pitches and dimensions, are analyzed in this Subsection. In all cases, a stack of two $8 \times 8mm^2$ dies in a F2F configuration is considered. A total of 12.8W is dissipated on top of the bottom die, according to the power map shown in Figure 9.7, while no power is dissipated on the top die. The μbump layout is based on the JEDEC standard for the Wide-IO configuration [45]. This standard defines the features and the layout of the interconnections between logic and memory chips and it consists of four rectangular μbump arrays in the center of the chips. Additional thermal μbumps are added in correspondence to the heat dissipation regions but they cover a slightly smaller area than the high power region (cf. Figure 9.7). The modeled geometry is also sketched in Figure 9.7. The power map, the dimensions, the boundary conditions and the layout of the μbump arrays are kept constant in the three analyzed cases. The material properties of the interface material and of the μbumps arrays are, instead, defined for each single case in order to analyze their thermal impact. Table 9.3 lists the parameters used in the three models.

The top and bottom temperatures on the diagonal cross section for the three different cases are shown on the left hand side of Figure 9.8; *Case1* is represented

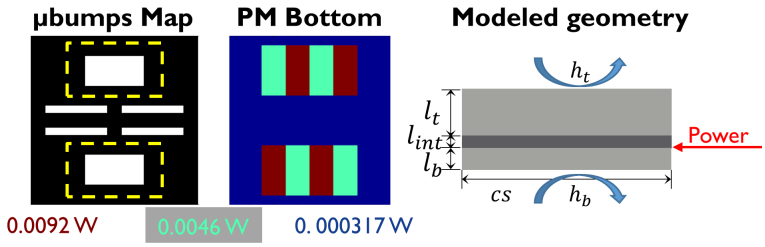


Figure 9.7: μBumps map, bottom power map and modeled geometry considered in the study of the interface material thermal impact. Yellow lines in the first picture indicate the corresponding high power dissipation area.

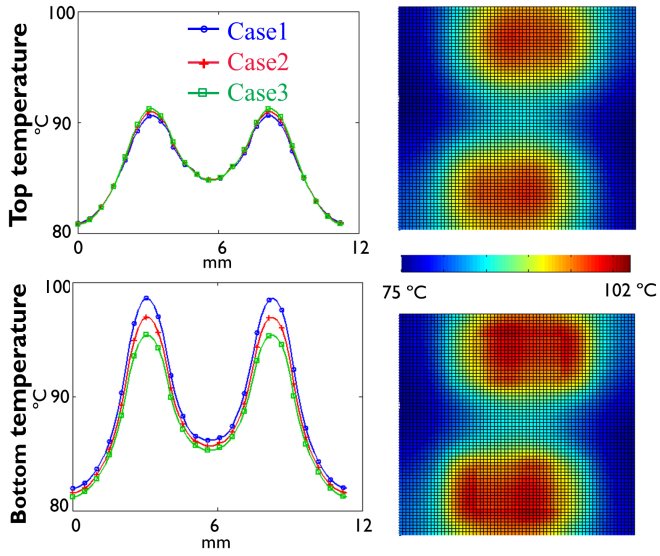


Figure 9.8: Left: comparison of the temperature profiles obtained for *Case1* (blue), *Case2* (red) and *Case3* (green) material on the diagonal of the top and bottom die. Right: complete temperature profiles on top and bottom die for *Case3* material.

by blue, *Case2* by red and *Case3* by green curves. The graphs clearly show the impact of the interface material on the temperature values. Since the power is dissipated on the bottom die and the heat is convectively removed from the top, the positive effect of using better conductive material is highly visible in the bottom temperature profile. From *Case1* to *Case3*, indeed, a 3.5% reduction on the maximum temperature increase is experienced. The use of better materials in the interface layer has, however, an impact also on the top temperature and, due to the improved interface thermal conductivity, from *Case1* to *Case3* this results in a slight increment of the top temperature (0.15% higher temperature increase). The

right column of the same graph shows the complete top and bottom temperature maps obtained with the materials selected for *Case3*.

Through this analysis, that can be easily performed changing some input numbers and recomputing the HSRs, it is possible to rapidly obtain information about the improvements in terms of temperature, and therefore, reliability, achievable by using better interface materials. This study can also help to decide whether, taking also the higher price for better material into account, the employment of more conductive material is beneficial or not.

9.4.2 Thermal impact of dummy μbumps

Another important early design phase question concerns the amount and the location of the dummy μbumps. They are included in the interface layer just to improve the thermal performances of the device and not for electrical connection purposes. For this reason, they can be positioned with a higher design freedom. However, the final price of the chip increases with the amount of included μstructures and, therefore, an accurate analysis of the involved costs versus performances has to be carried out. The FTM presented in this thesis is well suited for this analysis since it allows for the rapid and easy computation of the thermal impact of different μbump layouts. The study presented in this Subsection has been published in [58].

This study relates the μbump area ratio, $\tilde{\rho}$, to the temperature in the middle of the heat dissipation area. This is, in most cases, the location of the maximum temperature. Figure 9.9 shows the considered μbumps maps. The basic layout is constituted by the commercial Wide-IO configuration [45]. Rectangular arrays of increasing size, centered with the high power dissipation areas, are added. The basic cases of homogeneous underfill and μbump arrays materials are also considered. The last plot in the same figure shows the bottom PM; the red cross indicates the location to which the data in Figure 9.10 refer. No power is dissipated on the top die while the values of the thermal conductivities and of the other system parameters are the ones used in *Case1* in Section 9.4.1 (cf. Table 9.3). Also the geometry is the same as the one considered in Section 9.4.1 and sketched in Figure 9.7.

The results of this study are presented in Figure 9.10, where the bottom temperature in the defined location is shown as a function of the μbump area ratio, $\tilde{\rho}$. Red circles refer to FTM results while blue crosses to FEM ones. Some significant cases are highlighted. The graph shows the positive trend in temperature reduction gained by introducing extra dummy μbumps centered above the heat dissipation area. The thermal effect of the Wide-IO layout is minimal since its location is not aligned with the heat sources. As soon as some μbumps are placed aligned with the hot region, the temperature starts dropping. The temperature reduction trend, however, saturates after a while: when $\tilde{\rho}$ reaches the value of 0.3-0.4,

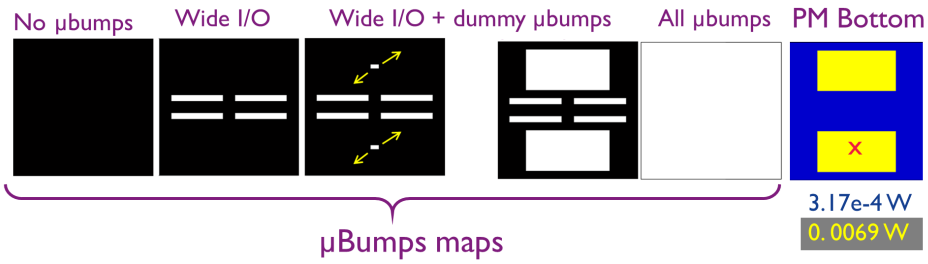


Figure 9.9: μBump layouts used in the study of the μbumps thermal impact (Figure 9.10). The last plot shows the PM used on the bottom die and the red cross indicates the position to which the data in Figure 9.10 refer. No power is dissipated on the top die.

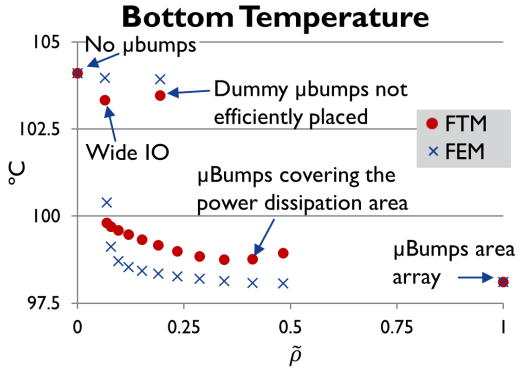


Figure 9.10: Bottom die temperature, in the location indicated in Figure 9.9, as a function of the μbump area ratio $\tilde{\rho}$. Red circles refer to FTM results while blue crosses to FEM ones. Significant cases are highlighted.

which corresponds to the coverage of the whole heat generating area, no extra improvement is achieved. The addition of extra dummy μbumps aligned with the hot regions on top of the Wide-IO configuration, up to covering 30% of the total interface area, results in a 7.5% temperature reduction for the bottom die.

Not only the amount of dummy μbumps is important, but also the location: the point $(\tilde{\rho}, T) = (0.2, 103°C)$, for example, is achieved by considering a single horizontal rectangular μbump array in the center of the interface layer. Since the array is not aligned with the hot areas, its effect is much smaller than the one obtained with the same $\tilde{\rho}$ value but placing the μbump arrays above the heat dissipation region. For the same amount of μbumps, a 30 times higher temperature reduction can, indeed, be achieved by placing them properly.

Comparison with FEM results shows a maximum error for these data of 1.3%. This

is really low but it has to be taken into account if different configurations want to be compared. Conclusions concerning if a specific design is better than another one can be drawn only if the difference is higher than 2-3%.

It should be noted that this analysis, to assess the best amount and placement of dummy μbumps in order to fulfill certain technological requirements while keeping the fabrication costs low, can be quickly and easily performed by just changing entries in the binary μbumps map. Once this input operation has been performed, less than 10 seconds are needed to obtain the temperature profiles for the specific configuration.

9.5 Summary

In this Chapter, possible applications of the FTM have been presented. More precisely, since for newly developed technologies thinned dies are used, the applicability of the FTM to study the thermal effect of extreme die thinning has been shown. Moreover, the thermal performance of a realistic power map (OpenSparc) has been analyzed considering different pulse trains in case of a 3D configuration and by comparing the thermal performances of a 3D and a 2D technology option. Finally, the FTM has been used to analyzed the thermal impact of different interface material properties as well as different layouts and amount of μbump arrays.

In all these cases, the application of the FTM is able to quickly and accurately provide an estimation of the temperature increase. It allows, therefore, to speed up and to simplify the comparison between different possible design options. Moreover, one of the main characteristics of this FTM is its ease of use even without having specific modeling expertise. The impact of different materials, dimensions, cooling strategies, μbump layouts and power distribution can be tested and studied by just modifying input numbers, entries in matrices and/or by computing new HSRs.

Chapter 10

General Conclusions and Recommendations

In this Chapter, an overview of the main findings and conclusions of this thesis are presented. The proposed algorithm can, at the present stage of development, certainly be used for the fast thermal modeling of microelectronic devices. However, further improvements and extensions are possible. Recommendations for further research or possible applications of the developed fast thermal model (FTM) are presented in Section 10.3.

10.1 Summary

Accurate thermal analysis of integrated circuits is crucial in the design phase of microelectronic devices. High temperature and temperature gradients may, indeed, activate a series of degradation mechanisms that ultimately lead the device to failure. These phenomena are even more pronounced in case of 3D-ICs since, in this technology, the active layers are placed on top of each other and the area available for cooling doesn't scale with the dissipated power. In order to produce reliable devices, accurate thermal modeling has to be performed, possibly already during the early design phase. This is normally accomplished by FEM models. If, on the one hand, this commonly accepted modeling methodology is able to provide accurate results, it is, on the other hand, computationally expensive and particular expertise is needed to properly develop the models. With the main aim of reducing computational time, in order to allow a quick comparison between different possible designs, various FTM methodologies have been presented in literature. As a drawback of the gained *speed-up*, they normally introduce a reduction in *accuracy*.

The work presented in this thesis is based on the semi-analytical Green's function approach, originally presented as *"Power Blurring"* in [37, 46, 83–86, 116–118], and on the *superposition work* published in [105]. It can be seen as a multiscale approach whose core allows the computation, at die stack level, of highly resolved temperature responses originated by any dissipated power map. The application of specific correction procedures on top of these results allows to include the temperature dependency of the silicon thermal conductivity as well as, on a coarser level of accuracy, the impact of the package thermal spreading and capacitance.

The core of the algorithm works by **convolving** power maps (PMs) and hot spot responses (HSRs) matrices. The first ones store the data concerning the location and the amount of the dissipated power while the second ones contain the information concerning the temperature responses of the system to localized, hot spot, power dissipation. They account, in particular, for the specific analyzed structure (geometry, material properties and boundary conditions) and they are computed by running easy, 2D-axisymmetric, FEM models for the stack configuration. In case of transient simulations, in particular, impulsive power dissipation is considered. Both the spatial and the temporal resolution of the resulting temperature profiles are uniform and user defined: the higher the resolution, the higher the computational time.

The superposition principle is applied to account for the thermal impact of power dissipated on different stacked dies while the method of images is used to account for the finite lateral dimension of the stack. In steady state, both the PMs and the HSRs are 2D matrices and 2D-convolution between them is performed. In the transient regime, instead, the time dependency has to be taken into account and, as a consequence, the matrices and the convolution operations become 3D. These results are presented in Chapters 2 and 3. The main original contribution of this work, in this initial phase, is the implementation of the 3D-convolution algorithm in the transient regime. In situations commonly encountered in thermal modeling of ICs, this algorithm allows a strong improvement in computational time when compared to the methodology, previously presented in literature, based on 2D spatial convolution plus time superposition.

Concerning the application of the correction procedure to include the **package thermal impact**, this is performed by comparing the temperature responses, for uniform power dissipation, computed considering two different geometries: the stack configuration, to which the basic convolution methodology is applied, and the package configuration, which is the more realistic one. The former profile can be computed by applying the *annulus method*, which takes as input the already computed HSRs. This is an original contribution of this work and it is presented in Section 3.4.1 for the steady state and in Section 3.6 for the transient regime. Concerning the latter profile, it is obtained by running a coarse, simplified FEM model for the package configuration and uniform power dissipation. The ratio between these quantities allows to account for the thermal spreading and for the thermal capacitance that are neglected in the geometry considered in

the convolution based algorithm for the stack configuration. This correction is independent of the specific power dissipated in the die stack and it is computed just once for each specific structure (geometry, material and boundary conditions).

In case of the transient regime, however, since the thermal impact of the package varies with time, a 2D-convolution plus time superposition approach has to be implemented to include the package correction. This means that computational time, with respect to the stack FTM based on 3D-convolution, increases but, at the same time, accuracy *may* highly improve. The thermal impact of the package structure depends, indeed, on the package itself and on the applied cooling solutions. For these reasons, specific metrics have been proposed, for both the steady state and the transient regime, to estimate the maximum possible improvement in accuracy achievable by applying the package correction to a specific structure. Moreover, it has been shown that, if the interest is just in the maximum temperature at each time step, the 3D-convolution approach plus a time-independent package correction can be a quicker alternative. These are original contributions of this work and they have been presented in Chapter 5.

The importance of including the **temperature dependency of the material properties** (mainly silicon) in the FTM has also been discussed in this thesis (cf. Chapter 6). In a previous work available in literature, an *iterative* method has been proposed to include this non-linear effect in a convolution based FTM. In this thesis, a *one-step* methodology based on the Kirchhoff transformation, which is an analytical transformation of the temperature profiles obtained by the convolution based FTM, has been presented. If, on the one hand, this operation does not increase computational time, it is, on the other hand, extremely important that the initial, fixed value of the silicon thermal conductivity is chosen appropriately, depending on the overall amount of dissipated power in the chip. The wrong selection of this value can, indeed, result in high errors. This means, in particular, that the HSRs are not only dependent on the geometrical structure, on the considered materials and on the applied boundary conditions but also on the total dissipated power. Even if, the application of the Kirchhoff transformation itself is significant only in case of large temperature differences within the dies, the computation of the HSRs with an appropriate fixed value of the silicon thermal conductivity, based on the average temperature increase, is always advisable. The algorithm to include the temperature dependency of the silicon thermal conductivity has been fully developed for the steady state regime, while some indications have been provided for transient simulations. In this regime, in particular, multiple HSRs need to be computed to be able to account for the dependency of the silicon thermal conductivity on the average temperature increase, which is time dependent since the total amount of dissipated power varies with time. Moreover, at the actual stage of development, the application in the transient regime of the Kirchhoff transformation, which would partially account for the dependency of the material properties on the spatial and temporal temperature variations, is advisable only if the considered time step is larger that the time constant of the system, i.e $\Delta t \geq \tau$.

In Chapter 4, the FTM based on convolution has been extended to include the thermal impact of **die-die interconnections** in case of two dies stacks (without package) in a F2F configuration in the steady state regime. The approach, previously proposed in literature to account for this effect in a convolution based FTM, required a scan over each considered cell in the grid and the application of superposition. The method presented in this thesis, instead, requires just two convolution operations for each [temperature response layer-power dissipation layer] combination. This means just a double amount of convolution operations with respect to the basic FTM, which considers a homogeneous interface material layer. It is, in particular, not necessary to scan all the cells in the discretized power maps. By performing a weighted average between the two temperature profiles corresponding to the same [temperature response layer-power dissipation layer] combination, the local and the global thermal impact of specific μbump layouts is taken into account. The weights are based on fitting functions in which the geometric parameters as well as the material properties and the μbumps amount and locations are taken into account. The added value of this Chapter of the thesis is, therefore, the possibility to include both the local and the global thermal impact of specific μbump layouts maintaining the core of a quick, accurate and highly resolved modeling methodology based on convolution. This can be useful, for example, in the early design phase to optimize the placement and the amount of dummy μbumps (cf. Section 9.4.2).

The FTM has also been successfully validated with respect to FEM and **experimental** results for a packaged, 3D-IC in both the steady state and the transient regime in Chapter 7. For both the low power and the high power package configurations considered in the steady state regime, the maximum percentage error on the peak temperature between FTM and experimental results is around 5-6%. Concerning the transient experimental results, the absolute error with respect to the FTM results after 1 second of chip activity is around 1.3°C/W, if a hot spot is dissipated in the center of the die, and around 2.3°C/W if it is located in the corners, where the temperature increase, as well as the thermal impact of the package, is higher.

The speed-up obtained by the convolution based algorithm is significant, with respect to a methodology based on superposition, if a highly resolved temperature map is required. It is, however, possible to obtain a strong reduction in computational time if the required temperature information is **limited to a specific set of points**, the ones, for example, in which the temperature peaks are expected. This is possible accounting for all the information that would be considered, in a highly resolved convolution based approach, to compute the temperature increase in those specific points. This means, in particular, that the computational time strongly reduces while the same accuracy is maintained. If this is of little impact in the steady state regime, it can be extremely useful in transient simulations, especially if one or more corrections need to be applied on top of the convolution results and, therefore, the 3D-convolution based FTM cannot be implemented. Moreover, it is important to stress that, for the commonly used FEM models, to

obtain the temperature in selected points, the model has to compute the results in all the nodes, different points cannot be decoupled while computing the solution. This original contribution of this thesis has been presented in Section 5.4.6.

In case of the transient regime, moreover, the 3D-convolution based FTM requires a constant time step. If the package correction is not considered, this time step can be taken as large as the greatest common divisor of the time intervals in which the PMs remain constant. If the package correction has to be applied, however, a limitation on the selection of the initial time step Δt has to be imposed. As explained in Section 8.3.2, this initial Δt has to be smaller than the time at which the spreading and the capacitive effect of the package become significant. Depending on the situation, it may be necessary to consider a very small initial time step, which would result in high computational time and memory issues. The possibility to use a **variable time step approach** has been discussed in Section 7.4.3. In the proposed methodology, the lengths of the time steps are selected by the user (it is not an automated procedure) and their variation has to satisfy a specific constraint. They have, indeed, to be periodically repeated, with a period equal to the greatest common divisor of the time during which the PM does not change. This variable time step algorithm strongly improves the run time and the applicability of the model not only in case a limitation on the time step is imposed, but also in case both the short and the long range temperature responses of the system have to be computed. This original contribution of this thesis has also been used in the experimental validation of the FTM.

Finally, in Chapter 8, the model originally presented for packaged 3D-IC has been extended to **different geometries** commonly available for microelectronic applications. An interposer (steady state) and a pyramidal (transient) configuration have been considered. In both cases, a larger structure (interposer or larger die) is present in the lower part of the geometry and appropriate adjustments of the package correction approach have been successfully implemented to model these situations. This is the first time that a highly resolved, convolution based method has been applied to these geometries, proving that the package correction approach can be useful in more general situations in which thermal spreading occurs.

10.2 General Conclusions

As a general conclusion we can state that the developed FTM satisfies all the success criteria defined in Section 1.4: the main goal of this thesis was, indeed, to obtain an *easy-to-use*, *highly resolved* modeling methodology able to *quickly* and *accurately* predict the steady state and transient thermal behavior of *3D-ICs* for user defined *power dissipation* scenarios. While for the first two characteristics a general indication can be provided, the other two are more difficult to quantify in a general way.

The application of the model requires a basic knowledge of FEM, since both the HSRs and the package corrections are computed by FEM. All the required FEM models are, however, simple, they can be quickly built and run, and they don't require a deep understanding of the finite element analysis. The easiness-of-use of the developed FTM to test and compare different scenarios has also been proved by the proposed applications in Chapter 9. The number of steps required to compare, using the FTM, the thermal impact of different values of a specific parameter depends on the considered parameter: for μbump and power dissipation layouts, just the entries in the corresponding matrices have to be changed; for material properties or dimensions in the die stack, as well as for a different amount of dissipated power, the 2D-axisymmetric FEM models for the HSRs need to be recomputed; for different package options, the coarse FEM models for the package corrections (eventually also the HSRs) need to be rerun. Moreover, it is important to stress that, if the proprietary information on the geometrical and material properties cannot be disclosed, the HSRs (not the parameters) can be provided to the thermal engineers and the model can still be run as a "black box" to compare different power dissipation scenarios, package options or μbump layouts.

Accuracy and computational time mainly depend on the spatial and temporal resolution and they go in opposite directions. High resolution is certainly possible and it results in higher accuracy but, at the same time, in higher computational time. It has to be noted that, when the computational time of FEM and FTM are compared, just the running time is considered. This is because the time needed to build the full FEM model and/or the FEM models for the HSRs and for the package corrections (which are required for the FTM) highly depends on the person who builds the models. For the cases considered in this thesis (two dies, packaged 3D-IC), for example, more than two orders of magnitude speed-up in computational time, with respect to conventional FEM, is achieved both in the steady state and in the transient regime. The error with respect to FEM is kept below 5% in stationary regime and below 3°C in dynamic simulations with time varying PMs. Moreover, the temperature profiles are computed just at selected levels (or selected points), avoiding the calculations in regions of low thermal interest. The thermal influence of the whole structure is, nevertheless, always kept into account. This is a great advantage with respect to FEM models, for which, even if the result is needed just in a limited number of points, the full 3D geometry needs to be discretized and the temperature has to be computed in all nodes. In this way, computational time is reduced and high accuracy is maintained.

The developed algorithm, whose actual stage of development is schematized in the flowcharts in Figures 6.6 and 6.11, is applicable to analyze the thermal behavior of real 3D-ICs. It has, indeed, been successfully validated with respect to experimental results and it accounts for the package thermal impact, the temperature dependency of the material properties and the possibility to considered N stacked dies. Even if, in the thesis, all the examples refer to two dies stack, the extension to N dies stacks is straightforward. The possibility to include the thermal impact of specific μbump layouts, at the present stage of development, is, however, restricted to

two dies stacks in the steady state regime. This algorithm is schematized in the flowchart in Figure 4.15 and further research needs to be performed in this area if more non-uniform layers have to be considered.

Throughout the thesis it has been proved that, for typical 3D-ICs applications, the basic limitation of the convolution based model, which requires the temperature responses of the system to be independent of the position where power is dissipated, can be overcome by the application of appropriate corrections. More innovative cooling solutions (as liquid cooling, intralayers cooling, jet impingement,...), which also violate this basic requirement, have not been considered in this thesis. However, it has been shown that approaches similar to the one considered for the thermal modeling of 3D-ICs can be successfully implemented to thermally analyze different geometries commonly available for microelectronic applications (interposer, pyramidal structure,...).

The developed FTM proved, therefore, to be a valid alternative to FEM in case of steady state and transient thermal simulations of 3D-ICs: the computational time is, indeed, strongly reduced and high accuracy is maintained.

10.3 Recommendations for further research

The presented FTM algorithm is a working methodology and it can already be applied to the thermal analysis of microelectronic devices. In the following of this Section, some indications are given on possible improvements in the implementation of the algorithm, as well as on further possible extensions and applications to a broader range of devices.

10.3.1 Implementation

Parallelization & GPU implementation A first step to be considered in the further development of the algorithm might be a simple improvement in the implementation of the code in order to fully exploit the parallelization and GPU capabilities of the computer on which the algorithm is run. Running the code in parallel might considerably reduce the computational time, especially in the transient regime if the 2D-convolution plus time superposition algorithm is applied. In this situation, indeed, the time dependent temperature responses of the system to the power dissipated at *each* specific time step are computed separately. Results corresponding to the same point in time are superposed afterwards. The 2D-convolution part of this algorithm is, therefore, run independently for each power scenario dissipated at a specific point in time and it can be easily parallelized. GPU computation can, instead, be used to improve the performance of the fast

Fourier transform algorithm used in the convolution operations. According to the technical documentation of Matlab, a five time speed-up can be expected [67].

Parametrization of the HSRs The HSRs are currently computed by 2D-axisymmetric FEM models. These models are simple to be created, they are quick to be run and they provide accurate thermal information over the specific stack that is considered. They need, however, to be built and run for each specific stack configuration that is considered. In this work, the parametrization of the HSRs has not been considered because the inaccuracy that would be introduced by applying this strategy, and that would spread and blow up by applying convolution, won't probably be compensated by the improvement in computational time. This step might, however, be necessary in further applications (as optimization, for example) and an attempt of implementing it was published in [83]. In that case, just the thermal conductivity of the substrate, the convection coefficient applied on one side of the model and the thickness of a single die have been considered as parameters in the fitting procedure.

The main issue in this frame would be, in my opinion, to derive a reliable fitting function that can account for the variation of all the parameters in an accurate way. A small error in the HSR can result in a huge error in the obtained temperature maps. The convolution operation is, indeed, highly sensitive to variations and noise in the convolved matrices.

10.3.2 Extensions

Active larger dies In Chapter 8, the FTM developed for packaged 3D-ICs in which all the dies have the same floorplan area has been extended to deal with the interposer and the pyramidal configurations. The transient implementation for the interposer configuration has not been considered in this thesis but I expect an approach analogous to the one implemented for the pyramidal structure to work. What is still missing in both the proposed extensions is the presence of an *active* larger die. If power is dissipated on the interposer itself or on the bottom larger die it might be possible, depending on the applied package and cooling solution, that the constriction resistance towards the smaller dies plays an important role in defining the final temperature profiles. Further investigation is needed in this case and, as a first attempt, an approach similar to the package correction one might be considered. The impact of the geometry might, however, be quite different depending if power is dissipated only in the region aligned with the smallest die, only in area not aligned with the smallest die or in both of them.

High power transistors The developed FTM can also be applied to high power and power switching applications. In GaN transistors, for example, higher current

densities are imposed on smaller volumes compared with silicon power devices leading, therefore, to higher power densities and higher temperatures [99]. For these reasons, the extension of the presented FTM to this new range of applications might be of interest for future research. The FTM would allow, indeed, to obtain a distributed temperature field, not just the junction temperature as it is commonly done at the moment. In these devices, power is generated in the power transistors, which can be simplified as high-aspect ratio rectangles (typically hundreds of μm length vs few μm width), and a distributed temperature model allows to account for the lateral heat spreading as well as for the maximum temperature. Moreover, the implementation of the FTM in this frame would allow to quickly evaluate the thermal impact of the layout design as, for example, the gate-to-gate spacing, the number of considered gates or the length of the gates. Due to the particular layout of the power map, it might be interesting to consider a rectangular grid instead of a square one. Moreover, since GaN transistors are normally used in power switching applications in which pulse trains are considered, care should be taken to include the different time constants of the system. The power transistors, indeed, thermally react in the range of nanoseconds but the package and the BCs have an impact in the range of 100 milliseconds or seconds. The selection of an appropriate time step and of a smart way to include the thermal impact of the slow materials is of utmost importance in this scenario.

10.3.3 Applications

Thermal aware design This might be an interesting application of the FTM, especially from a design point of view. Different kind of optimization studies can be considered.

The easiest one is the optimization of the **switching time** of specific power areas (or cores) for a fixed design of the power maps. This means that both the structure of the 3D-ICs (and, as a consequence, the HSRs and the package correction profiles) and the basic layout of the dissipated power maps are kept constant. The parameter to be optimized is either the maximum amount of power that can be dissipated in a specific time interval, or the length of the time interval in which a certain amount of power can be dissipated, before a specific core needs to be switched off or before the load needs to be moved to another core. In both cases, the maximum temperature has to be constraint within safe limits. These limits can be specifications on the maximum temperature allowed in the active components, such as logic and memory dies, or on the temperature allowed on the case of the system if, for example, handheld devices are considered. Moreover, it would be interesting to include also the possibility to consider over-clocking of the cores, meaning that they run at higher speed than their nominal value.

Another possible optimization option concerns the **dimensions and the material properties of the stack**. The steady state regime is probably the easiest and most

interesting scenario to be considered. In this case, the power map is kept constant while the BCs, the thicknesses and the materials of the various layers in the stack can vary. To this aim, the parametrization of the HSRs is required. If a fixed package option is considered (LP or HP), it is not to be expected that the package thermal impact changes significantly due to variation of the die stack itself and, as a consequence, there is no need to parametrize the package correction. If, however, also the thickness and the materials of the different layers representing the package in the HSRs are considered as variable in the optimization, it might be necessary to adapt the package correction profile accordingly.

Finally, a more complex optimization process might be the one concerning the thermal optimization of the **floorplan of the power maps**. In this case, a specific structure for the 3D-ICs and a certain amount of power blocks to be placed on the PM itself are given. Also this scenario should be treated in the steady state regime and a weighted average of the maximum temperature and of the temperature gradients should be minimized. If, in particular, the footprint area and the total dissipated power are kept constant, the cost function can include just the maximum temperature. This is because, the minimization of this quantity automatically results in the minimization of the spatial temperature gradients. However, in order to improve the system from a thermal point of view, it is advisable to leave a certain freedom to the footprint area of the device. Examples of temperature aware floorplanning algorithms are presented in literature and they are based, for example, on simulated annealing [25] and on a forced-directed approach [115]. They all consider a resistance based FTM for the thermal simulations. It would be interesting to apply these, or similar, optimization methodologies coupled with the highly resolved convolution based FTM to further optimize the placements of the blocks. It might be useful, furthermore, to take advantage of the possibility of the developed FTM to compute the temperature just in specific locations (centers of the power areas, for example) maintaining the same high accuracy of the highly resolved FTM. While running the optimization it is, however, necessary to take into account both the electro-thermal coupled phenomena and the electrical constraints that are required to have a reliable and functional device from that point of view. Some blocks need, indeed, to be placed close to each other and, due to self heating, the relocation of some power areas, and the consequent different length of the interconnect lines, might also influence the amount of power that is dissipated.

Appendix A

FEM models

In the thesis, different FEM models have been used to validate the developed FTM. Most of them are simplifications or adaptations of a single FEM model, originally developed by dr. ir. Herman Oprins in Marc MSC for the PTCQ test chip [73–75]. In this Appendix, a description of the model is presented together with a grid convergence analysis.

FEM results have been also used to build the FTM itself (HSRs and package correction). A description of the FEM model used to compute the HSRs has been reported in Section 2.5.1 while, in this appendix, a more accurate evaluation of the accuracy of the coarse model, used to generate the package corrections and introduced in Section 5.3.4, is presented.

A.1 FEM model for the PTCQ at package level

In this Section, the FEM model for the PTCQ test chip in a low power package configuration is presented. The layout of the test chip and the schematic of the package cross section are shown in Figure 7.1 and Figure 7.3, respectively. Figure A.1 presents the geometry and the mesh used for the FEM model. The layouts of the die-die interface layer and of the layer connecting the bottom die to the package substrate are also shown. Just 1/4 of the geometry, which is symmetrically repeated in the missing directions, is illustrated in the Figure. However, to be able to apply a non-symmetric PM, the full model has been built and run. The dimensions of each layer and the thermal properties of each material are reported in Table 7.1 in Section 7.4. Just the underfill material ($k = 0.4 W/mK$) is missing in that Table because the listed parameters referred to the FTM, in which uniform material layers are assumed.

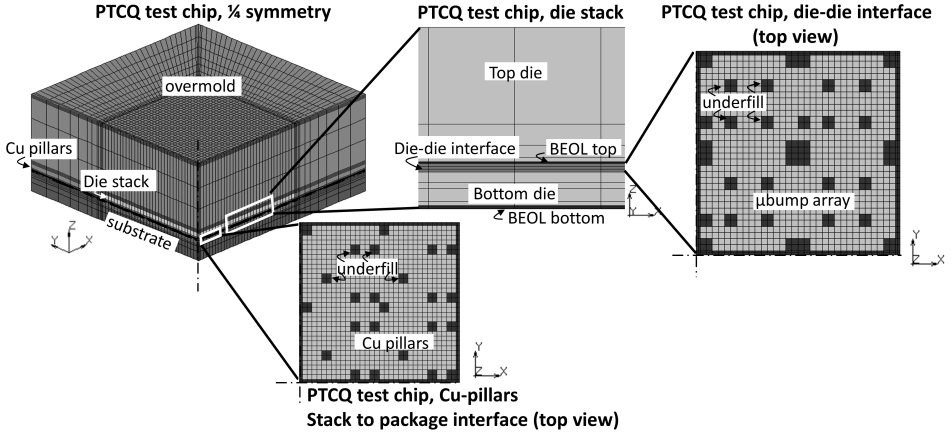


Figure A.1: Geometry and mesh used for the FEM model of the PTCQ test chip in LP configuration; 1/4 symmetry. The layouts of the μbump and Cu-pillar arrays are also indicated.

μBumps and Cu-pillars are not modeled individually but equivalent material properties for the arrays are extracted. The locations named in Figure A.1 as *Cu pillars* and *µbump array* refer, therefore, to areas in which, respectively, both Cu pillars and underfill or μbumps and underfill are present. In the cells indicated by *underfill*, instead, just underfill material is present and modeled. This is done according to the floorplan shown in Figure 7.2. The equivalent material properties assigned to Cu-pillars and to μbump arrays are computed using a unit cell modeling approach. By using this approach, 1/4 of a single μbump is modeled together with the underfill surrounding it and covering a distance of 1/2 pitch. In this way, both the dimensions of the single μbump (diameter and thickness) and the distance between consecutive μbumps in the array are taken into account. By applying appropriate boundary conditions and heat fluxes, both the in-plane and the out-of-plane equivalent thermal conductivities can be computed [73–75].

Comparing Figure A.1 and Figure 7.1 in Section 7.2, which shows the arrangement of the different functional cells in the PTCQ, it is clear that each functional cell is represented by four elements in the $x-y$ plane. The horizontal dimensions of these elements are, therefore, $120{\times}120\mu m^2$ and the full model consists of approximately 250000 elements. This kind of mesh allows to model every PM that can be experimentally applied to the test chip. To account for the temperature dependency of the silicon thermal conductivity, three subsequent iterations are run in each simulation.

Concerning the accuracy of the model, a *grid convergence* analysis has been performed [48, 92, 93]. By running three FEM models with subsequent grid refinements, it is, indeed, possible to estimate an error bound for the provided result. To this aim, first of all, the order p of grid convergence has to be calculated.

This is possible by computing three subsequent solutions. More precisely, these solutions (f_3, f_2, f_1) are computed over three different grid levels ($\bar{h}_3, \bar{h}_2, \bar{h}_1$), which are subsequently refined according to a constant grid refinement ratio r, i.e. $\bar{h}_1 = \frac{\bar{h}_2}{r} = \frac{\bar{h}_3}{r^2}$. Using this notation,

$$p = \frac{\ln\left(\frac{f_3 - f_2}{f_2 - f_1}\right)}{\ln r}. \tag{A.1}$$

Once the order of convergence is known, the grid convergence index (GCI), which provides an error bound on the computed solution, can be calculated by using two subsequent results. In particular, if f_3 and f_2 are used and the final reported result is f_3, the one on the coarsest grid,

$$GCI = \frac{F_s r^p}{r^p - 1}\left|\frac{f_3 - f_2}{f_2}\right| \tag{A.2}$$

where F_s is a safety factor. In particular, $F_s = 1.25$ in case three grid levels are considered (as in this case) while $F_s = 3$ if just two grid levels are taken into account. It is also important to be sure that the selected grid levels are in the asymptotic range of convergence for the computed solution. This can be done by checking if

$$\frac{GCI_{23}}{r^p GCI_{12}} \approx 1, \tag{A.3}$$

where GCI_{23} and GCI_{12} are the values of GCI computed by considering, respectively, f_2, f_3 and f_1, f_2.

In this analyzed case, $r = 2$, $\bar{h}_3 = 120\mu m$, $\bar{h}_2 = 60\mu m$ and $\bar{h}_1 = 30\mu m$ for the elements corresponding to the dies stack. The results reported in the thesis are the ones corresponding to f_3 in the subsequent mesh refinement. This is because, since a single model has been used to test different power dissipation scenarios (single HS, uniform power, arbitrary PMs), an a priori localized refinement of the mesh is not possible and the use of the fine mesh considered to obtain f_1 strongly increases the complexity of the model and the computational time. The results reported hereafter refer to the PM shown in Figure 7.6, which was considered for the experimental study, and to uniform power dissipation. Due to the limitation of the hardware, the analysis (and the mesh refinement) has been limited to a critical area of the chip, the one in which the maximum temperature is expected (for the specific PM this is the HS in the bottom right corner). The maximum temperature values for both power dissipation scenarios and for the three different levels of discretization are listed, for both the top and the bottom die, in Table A.1 while the values obtained for the *GCI* in the same locations are reported in Table A.2.

For both PMs, the grid levels are well within the asymptotic range of convergence for the computed solution. The *GCI* values for the specific PM with multiple HSs are quite high, while the ones obtained for uniform power dissipation

Table A.1: Maximum temperatures obtained, in the grid refinement analysis, for the specific PM shown in Figure 7.6 and for uniform power dissipation.

	PM		Uniform	
Grid size	max(ΔT) top die	max(ΔT) bottom die	max(ΔT) top die	max(ΔT) bottom die
$\bar{h}_3(120\mu m)$	43.7 °C	44.5°C	64.5°C	64.7°C
$\bar{h}_2(60\mu m)$	44.1 °C	45°C	64.6°C	64.7°C
$\bar{h}_1(30\mu m)$	44.5°C	45.3°C	64.6°C	64.8°C

Table A.2: *GCI* values and check for the asymptotic range of convergence in case of FEM results for the LP PTCQ package; specific PM shown in Figure 7.6 and uniform power dissipation.

	PM		Uniform	
	GCI	convergence	*GCI*	convergence
top die	3.9%	1.007	0.86%	1.002
bottom die	4%	1.007	0.92%	1.002

are well acceptable. The model with grid size $\bar{h}_3$ consists of 250000 elements and, considering three iterations to account for the non-linearity introduced by the temperature dependency of the silicon thermal conductivity, it takes about two minutes to run (Marc MSC run, without parallelization license, on a HP Proliant DL360pGen8 server with 16 Intel(R) Xeon(R) CPU E5-2650 v2 @ 2.60GHz). Considering $\bar{h}_1$ as mesh size results in a model with 16 millions elements, which becomes intractable with the available hardware and software. For these reasons, and since a uniform mesh is needed for the die stack due to the requirement of having a single model to test different PMs, a compromise on the grid resolution has to be made and the coarser grid, with still an acceptable discretization error, has been chosen.

This LP PTCQ model represents the base for almost all the FEM models used in this thesis. The HP PTCQ model, for example, is very similar to this one, only the overmold is removed and appropriate BCs are directly applied on top of the top die. The model used for the interposer geometry in Section 8.2 is made of two PTCQ models placed next to each other with a proper package representation. Also the FEM model used to validate the package correction approach in Section 5.3.6 is a simplified version of the LP PTCQ model in which an area array of μbumps is considered. Due to the similarity of all these models, just the grid convergence analysis for the PTCQ in the LP package configuration is reported.

Table A.3: *GCI* values and check for the asymptotic range of convergence in case of the coarse FEM model for uniform power dissipation in case of LP and HP package configurations.

	LP		HP	
	GCI	convergence	*GCI*	convergence
max(T)	0.25%	1.0005	0.0029%	1
min(T)	0.47%	1.0009	0.0996%	1.0002

A.2 Coarse FEM model for the package

In Section 5.3.4 the package correction approach has been presented for the steady state regime. A solution based on a coarse FEM model simulation for uniform power dissipation has been proposed in order to include the package thermal impact in the FTM. The mesh and the geometry used for a LP package configuration are shown in Figure 5.11 (b). The geometry considered for the HP package is similar to the presented one, just the overmold is removed and the BCs are directly applied on top of the top die. Due the importance of these coarse models in the FTM, a grid convergence analysis has been run and the results, for both the LP and the HP packages, are reported in Table A.3. Moreover, since just the maximum and minimum temperature values are extracted from these coarse models, the grid convergence analysis has been run for these extremes.

As Table A.3 shows, in all cases the error bounds on both the maximum and the minimum extracted temperature values are small and the grid levels are well within the asymptotic range of convergence for the computed solution. We can, therefore, conclude that grid independence is achieved for the coarse models for both the LP and the HP package configurations and, due to the main requirement of these coarse models to be run fast, in the thesis the coarser grid has been considered.

Bibliography

[1] Astrid, P. *Reduction of process simulation models: a proper orthogonal decomposition approach*. Technische Universiteit Eindhoven, 2004.

[2] Augustin, A., and Hauck, T. A new approach to boundary condition independent compact dynamic thermal models. In *Semiconductor Thermal Measurement and Management Symposium, 2007. SEMI-THERM 2007. Twenty-Third Annual IEEE* (2007), IEEE, pp. 228–232.

[3] Bagnall, K. R., Muzychka, Y. S., and Wang, E. N. Application of the Kirchhoff transform to thermal spreading problems with convection boundary conditions. *Components, Packaging and Manufacturing Technology, IEEE Transactions on 4*, 3 (2014), 408–420.

[4] Barabadi, B., Joshi, Y. K., and Kumar, S. Rapid multi-scale transient thermal modeling of packaged microprocessors using hybrid approach. In *Electronics Packaging Technology Conference (EPTC), 2012 IEEE 14th* (2012), IEEE, pp. 157–164.

[5] Batty, W., David, S., and Snowden, C. Reply to comment on 'Electro-thermal device and circuit simulation with thermal nonlinearity due to temperature dependent diffusivity'. *Electronics Letters 37*, 24 (2001), 1482–1483.

[6] Batty, W., and Snowden, C. Electro-thermal device and circuit simulation with thermal nonlinearity due to temperature dependent diffusivity. *Electronics Letters 36*, 23 (2000), 1966–1968.

[7] Bejan, A., and Errera, M. R. Convective trees of fluid channels for volumetric cooling. *International Journal of Heat and Mass Transfer 43*, 17 (2000), 3105–3118.

[8] Beneventi, F., Bartolini, A., Tilli, A., and Benini, L. An effective gray-box identification procedure for multicore thermal modeling. *Computers, IEEE Transactions on 63*, 5 (2014), 1097–1110.

[9] Beyne, E. 3D system integration technologies. In *VLSI Technology, Systems, and Applications, 2006 International Symposium on* (2006), IEEE, pp. 1–9.

[10] Bonani, F., and Ghione, G. On the application of the Kirchhoff transformation to the steady-state thermal analysis of semiconductor devices with temperature-dependent and piecewise inhomogeneous thermal conductivity. *Solid-state electronics 38*, 7 (1995), 1409–1412.

[11] Briggs, W. L., and Henson, V. E. *The DFT: An Owner's Manual for the Discrete Fourier Transform*. SIAM, 1995.

[12] Brunschwiler, T., Michel, B., Rothuizen, H., Kloter, U., Wunderle, B., Oppermann, H., and Reichl, H. Forced convective interlayer cooling in vertically integrated packages. In *Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2008. 11th IEEE Intersociety Conference on* (2008), IEEE, pp. 1114–1125.

[13] Brunschwiler, T., Schindler-Saefkow, F., Gordin, R., Haupt, M., and Schlottig, G. Study of the compound of properties of percolating and neck-based thermal underfills. In *Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2014. 14th IEEE Intersociety Conference on* (2014), IEEE, pp. 227–234.

[14] Chanchani, R. *Materials for Advanced Packaging*. Springer US, Boston, MA, 2009, ch. 3D Integration Technologies – An Overview, pp. 1–50.

[15] Cherman, V., Lofrano, M., Simons, V., Gonzalez, M., Van der Plas, G., De Vos, J., Wang, T., Daily, R., Salahouelhadj, A., Beyer, G., et al. Effects of packaging on mechanical stress in 3D-ICs. In *Electronic Components and Technology Conference (ECTC), 2015 IEEE 65th* (2015), IEEE, pp. 354–361.

[16] Cherman, V., Van der Plas, G., De Vos, J., Ivankovic, A., Lofrano, M., Simons, V., Gonzalez, M., Vanstreels, K., Wang, T., Daily, R., et al. 3D stacking induced mechanical stress effects. In *Electronic Components and Technology Conference (ECTC), 2014 IEEE 64th* (2014), IEEE, pp. 309–315.

[17] Choobineh, L., and Jain, A. Analytical solution for steady-state and transient temperature fields in vertically stacked 3-D integrated circuits. *Components, Packaging and Manufacturing Technology, IEEE Transactions on 2*, 12 (2012), 2031–2039.

[18] Choudhury, A., Kothari, S., Mahanta, N., Dhavaleswarapu, H., and Chang, J. Compact thermal modeling methodology for active and thermal bumps in 3D microelectronic packages. In *ASME 2015 International Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems collocated with the ASME 2015 13th International Conference on Nanochannels, Microchannels, and Minichannels* (2015), American Society of Mechanical Engineers.

[19] Christiaens, F., Vandevelde, B., Beyne, E., Mertens, R., and Berghmans, J. A generic methodology for deriving compact dynamic thermal models,

applied to the PSGA package. *Components, Packaging, and Manufacturing Technology, Part A, IEEE Transactions on 21*, 4 (1998), 565–576.

[20] Codecasa, L. Canonical forms of one-port passive distributed thermal networks. *Components and Packaging Technologies, IEEE Transactions on 28*, 1 (2005), 5–13.

[21] Codecasa, L., D'amore, D., and Maffezzoni, P. A novel approach for generating boundary condition independent compact dynamic thermal networks of packages. In *IEEE Computer Society* (2004), pp. 305–310.

[22] Codecasa, L., D'Amore, D., and Maffezzoni, P. Boundary condition independent compact models of dynamic thermal networks with many heat sources. In *Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2006. 10th IEEE Intersociety Conference on* (2006), pp. 685–689.

[23] Codecasa, L., D'Amore, D., and Maffezzoni, P. Canonical forms of multiport dynamic thermal networks. In *Thermal Investigation of ICs and Systems, 2006. THERMINIC 2006. 12th International Workshop on* (2006), TIMA Editions, France, pp. 59–64.

[24] Cole, K. D., Beck, J. V., Haji-Sheikh, A., and Litkouhi, B. *Heat conduction using Green's functions*. Taylor and Francis, 1991.

[25] Cong, J., Wei, J., and Zhang, Y. A thermal-driven floorplanning algorithm for 3D ICs. In *Computer Aided Design, 2004. ICCAD-2004. IEEE/ACM International Conference on* (2004), IEEE, pp. 306–313.

[26] Coşkun, A. K., Whisnant, K. A., Gross, K. C., et al. Static and dynamic temperature-aware scheduling for multiprocessor SoCs. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on 16*, 9 (2008), 1127–1140.

[27] Culham, J. R., Yovanovich, M. M., and Lemczyk, T. Thermal characterization of electronic packages using a three-dimensional Fourier series solution. *Journal of Electronic Packaging 122*, 3 (2000), 233–239.

[28] De Munck, K., Chiarella, T., De Moor, P., Swinnen, B., and Van Hoof, C. Influence of extreme thinning on 130-nm standard CMOS devices for 3-D integration. *Electron Device Letters, IEEE 29*, 4 (2008), 322–324.

[29] De Vos, J., Jourdain, A., Erismis, M., Zhang, W., De Munck, K., Manna, A. L., Tezcan, D., and Soussan, P. High density 20μm pitch CuSn microbump process for high-end 3D applications. In *Electronic Components and Technology Conference (ECTC), 2011 IEEE 61st* (2011), IEEE, pp. 27–31.

[30] Ellison, G. N. Maximum thermal spreading resistance for rectangular sources and plates with nonunity aspect ratios. *Components and Packaging Technologies, IEEE Transactions on 26*, 2 (2003), 439–454.

[31] Fish, M., McCluskey, P., and Bar-Cohen, A. Modeling thermal spreading resistance in via arrays. In *ASME 2015 International Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems collocated with the ASME 2015 13th International Conference on Nanochannels, Microchannels, and Minichannels* (2015), American Society of Mechanical Engineers.

[32] F.M., W. *Fluid Mechanics*. McGraw–Hill, 2003.

[33] Gerstenmaier, Y., Kiffe, W., and Wachutka, G. Combination of thermal subsystems modeled by rapid circuit transformation. In *Thermal Investigation of ICs and Systems, 2007. THERMINIC 2007. 13th International Workshop on* (2007), IEEE, pp. 115–120.

[34] Gerstenmaier, Y., and Wachutka, G. Time dependent temperature fields calculated using eigenfunctions and eigenvalues of the heat conduction equation. *Microelectronics journal 32*, 10 (2001), 801–808.

[35] Gerstenmaier, Y., and Wachutka, G. Rigorous model and network for transient thermal problems. *Microelectronics journal 33*, 9 (2002), 719–725.

[36] Gerstenmaier, Y. C., and Wachutka, G. K. Transient temperature fields with general nonlinear boundary conditions in electronic systems. *Components and Packaging Technologies, IEEE Transactions on 28*, 1 (2005), 23–33.

[37] Heriz, V. M., Park, J.-H., Kemper, T., Kang, S.-M., and Shakouri, A. Method of images for the fast calculation of temperature distributions in packaged VLSI chips. In *Thermal Investigation of ICs and Systems, 2007. THERMINIC 2007. 13th International Workshop on* (2007), pp. 18–25.

[38] Huang, W., Ghosh, S., Velusamy, S., Sankaranarayanan, K., Skadron, K., and Stan, M. R. HotSpot: A compact thermal modeling methodology for early-stage VLSI design. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on 14*, 5 (2006), 501–513.

[39] Huang, W., Stan, M. R., and Skadron, K. Physically-based compact thermal modeling – achieving parametrization and boundary condition independence. In *Proc. 10th Int. Workshop THERMal Investigations of ICs Syst* (2004), pp. 287–292.

[40] Huang, W., Stan, M. R., Skadron, K., Sankaranarayanan, K., Ghosh, S., and Velusamy, S. Compact thermal modeling for temperature aware design. In *Proceedings of the 41st annual Design Automation Conference* (2004), ACM, pp. 878–883.

[41] Jain, A., Jones, R. E., Chatterjee, R., and Pozder, S. Analytical and numerical modeling of the thermal performance of three-dimensional integrated circuits. *Components and Packaging Technologies, IEEE Transactions on 33*, 1 (2010), 56–63.

[42] James, C. Application of conformal mapping and variational method to the study of heat conduction in polygonal plates with temperature/dependent conductivity. *International Journal of Heat and Mass Transfer 14*, 1 (1971), 49–56.

[43] Janicki, M., De Mey, G., and Napieralski, A. Transient thermal analysis of multilayered structures using Green's functions. *Microelectronics Reliability 42*, 7 (2002), 1059–1064.

[44] Janicki, M., Torzewicz, T., Vass-Varnai, A., and Napieralski, A. Impact of nonlinearities in boundary conditions on device compact thermal models. In *Thermal Investigation of ICs and Systems, 2013. THERMINIC 2013. 19th International Workshop on* (2013), IEEE, pp. 202–205.

[45] JESD229. Wide I/O Single Data Rate (Wide I/O SDR), 2011. `http://www.jedec.org/standards-documents/results/jesd229`.

[46] Kemper, T., Zhang, Y., Bian, Z., and Shakouri, A. Ultrafast temperature profile calculation in IC chips. In *Thermal Investigation of ICs and Systems, 2006. THERMINIC 2006. 12th International Workshop on* (2006).

[47] Krabbenhoft, K., and Damkilde, L. Comments on 'Electro-thermal device and circuit simulation with thermal nonlinearity due to temperature dependent diffusivity'. *Electronics Letters 37*, 24 (2001), 1481–1482.

[48] Kwaśniewski, L. Application of grid convergence index in FE computation. *Bulletin of the Polish Academy of Sciences: Technical Sciences 61*, 1 (2013), 123–128.

[49] La Manna, A., Rebibis, K. J., De Vos, J., Bogaerts, L., Gerets, C., and Beyne, E. Small pitch micro-bumping and experimental investigation for under filling 3D stacking. In *IMAPS, 45th International Symposium on Microelectronics* (2012), pp. 535–541.

[50] Lasance, C. J. Ten years of boundary-condition-independent compact thermal modeling of electronic parts: A review. *Heat Transfer Engineering* (2008).

[51] Lasance, C. J. M. Thermally driven reliability issues in microelectronic system: Status-quo and challenges. *Microelectronics Reliability 43*, 12 (2003), 1969–1974.

[52] Lee, S., Song, S., Au, V., and Moran, K. P. Constriction/spreading resistance model for electronics packaging. In *Proceedings of the 4th ASME/JSME thermal engineering joint conference* (1995), vol. 4, pp. 199–206.

[53] Li, P., Pileggi, L. T., Asheghi, M., and Chandra, R. Efficient full-chip thermal modeling and analysis. In *Computer Aided Design, 2004. ICCAD-2004. IEEE/ACM International Conference on* (2004), IEEE, pp. 319–326.

[54] Li, P., Pileggi, L. T., Asheghi, M., and Chandra, R. IC thermal simulation and modeling via efficient multigrid-based approaches. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on 25*, 9 (2006), 1763–1776.

[55] Liu, Z., Tan, S. X.-D., Wang, H., Hua, Y., and Gupta, A. Compact thermal modeling for packaged microprocessor design with practical power maps. *Integration, the VLSI journal 47*, 1 (2014), 71–85.

[56] Mackowski, D. W. Conduction heat transfer, notes for MECH 7210, 2015. `http://www.eng.auburn.edu/~dmckwski/mech7210/condbook.pdf`.

[57] Maggioni, F., Oprins, H., Beyne, E., De Wolf, I., and Baelmans, M. Convolution based compact thermal model for 3D-ICs: methodology and accuracy analysis. In *Thermal Investigation of ICs and Systems, 2013. THERMINIC 2013. 19th International Workshop on* (2013), IEEE, pp. 152–157.

[58] Maggioni, F., Oprins, H., Beyne, E., De Wolf, I., and Baelmans, M. Convolution based compact thermal model application to the evaluation of the thermal impact of die to die interface including interconnections. In *Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2014. 14th IEEE Intersociety Conference on* (2014), IEEE, pp. 98–106.

[59] Maggioni, F., Oprins, H., Beyne, E., De Wolf, I., and Baelmans, M. Convolution based steady state compact thermal model for 3D integrated circuits: methodology for including the thermal impact of die to die interconnections. In *Proceedings of the 15th International Heat Transfer Conference, IHTC-15 August 10-15, 2014, Kyoto, Japan* (2014), Begell House.

[60] Maggioni, F., Oprins, H., Beyne, E., De Wolf, I., and Baelmans, M. Fast convolution based thermal model for 3D-ICs: methodology, accuracy analysis and package impact. *Microelectronics Journal 45*, 12 (2014), 1746–1752.

[61] Maggioni, F., Oprins, H., Beyne, E., De Wolf, I., and Baelmans, M. Fast transient convolution based thermal modeling methodology for including the package thermal impact in 3D-ICs. *Components, Packaging, and Manufacturing Technology, IEEE Transactions on 6*, 3 (2016), 424–431.

[62] Maggioni, F. L. T., Oprins, H., Milojevic, D., Beyne, E., De Wolf, I., and Baelmans, M. 3D-Convolution based fast transient thermal model for 3D integrated circuits: methodology and applications. In *Thermal Measurement, Modeling & Management Symposium (SEMI-THERM), 2015 31st* (2015), IEEE, pp. 107–112.

[63] Masana, F. N. A closed form solution of junction to substrate thermal resistance in semiconductor chips. *Components, Packaging, and Manufacturing Technology, Part A, IEEE Transactions on 19*, 4 (1996), 539–545.

[64] Mathews, J. H., and Howell, R. W. Conformal mapping dictionary, 2008. http://mathfaculty.fullerton.edu/mathews/c2003/sourcesink/SourceSinkMod/Links/SourceSinkMod_lnk_5.html.

[65] Matlab, The MathWorks. fftn, 2015. http://nl.mathworks.com/help/matlab/ref/fftn.html.

[66] Matlab, The MathWorks. Matlab R2014a, 2014. http://www.mathworks.com/products/matlab/.

[67] Matlab, The MathWorks. Measure and improve GPU performance, 2015. http://www.mathworks.com/help/distcomp/measure-and-improve-gpu-performance.html.

[68] Moore, G. *Understanding Moore's Law: Four Decades of Innovation*. Chemical Heritage Foundation, 2006, ch. Chapter 7: Moore's law at 40, pp. 67–84.

[69] MSc Marc. Marc, 2015. http://www.mscsoftware.com/Products/CAE-Tools/Marc.aspx.

[70] Muzychka, Y., Culham, J., and Yovanovich, M. Thermal spreading resistance of eccentric heat sources on rectangular flux channels. *Journal of Electronic packaging 125*, 2 (2003), 178–185.

[71] Oprins, H., and Beyne, E. Generic thermal modeling study of the impact of 3D-interposer material and thickness options on the thermal performance and die-to-die thermal coupling. In *Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2014. 14th IEEE Intersociety Conference on* (2014), IEEE, pp. 72–78.

[72] Oprins, H., Cherman, V., Rebibis, K., Vermeersch, K., Gerets, C., Vandevelde, B., Beyer, G., Beyne, E., et al. Transient analysis based thermal characterization of die-die interfaces in 3D-ICs. In *Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2012. 13th IEEE Intersociety Conference on* (2012), IEEE, pp. 1395–1404.

[73] Oprins, H., Cherman, V., Van der Plas, G., De Vos, J., and Beyne, E. Experimental characterization of the vertical and lateral heat transfer in 3D-SiC packages. In *ASME 2015 International Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems collocated with the ASME 2015 13th International Conference on Nanochannels, Microchannels, and Minichannels* (2015), American Society of Mechanical Engineers.

[74] Oprins, H., Cherman, V., Van der Plas, G., Maggioni, F., De Vos, J., and Beyne, E. Thermal experimental and modeling analysis of high power 3D packages. In *IC Design & Technology (ICICDT), 2015 International Conference on* (2015), IEEE, pp. 1–4.

[75] Oprins, H., Cherman, V., Van der Plas, G., Maggioni, F., De Vos, J., Wang, T., Daily, R., and Beyne, E. Experimental thermal characterization and thermal model validation of 3D packages using a programmable thermal test chip. In *Electronic Components and Technology Conference (ECTC), 2015 IEEE 65th* (2015), IEEE, pp. 1134–1141.

[76] Oprins, H., Cherman, V., Vandevelde, B., Torregiani, C., Stucchi, M., Van der Plas, G., Marchal, P., and Beyne, E. Characterization of the thermal impact of Cu-Cu bonds achieved using TSVs on hot spot dissipation in 3D stacked ICs. In *Electronic Components and Technology Conference (ECTC), 2011 IEEE 61st* (2011), IEEE, pp. 861–868.

[77] Oprins, H., Cherman, V., Vandevelde, B., Van der Plas, G., Marchal, P., and Beyne, E. Numerical and experimental characterization of the thermal behavior of a packaged DRAM-on-logic stack. In *Electronic Components and Technology Conference (ECTC), 2012 IEEE 62nd* (2012), IEEE, pp. 1081–1088.

[78] Oprins, H., Maggioni, F., Vladimir, C., Van der Plas, G., and Beyne, E. *Handbook of 3D Integration – Volume 4: 3D Design, Test, and Thermal Management*. Wiley-VCH, 2016, ch. Thermal Modeling and Experimental Model Validation for 3D Stacked ICs.

[79] Oracle. OpenSparc T2, 2007. `http://www.oracle.com/technetwork/systems/opensparc/index.html`.

[80] Özdemir, I., Brekelmans, W., and Geers, M. Computational homogenization for heat conduction in heterogeneous solids. *International journal for numerical methods in engineering 73*, 2 (2008), 185–204.

[81] Palacin, J., Salleras, M., Samitier, J., and Marco, S. Dynamic compact thermal models with multiple power sources: Application to an ultrathin chip stacking technology. *Advanced Packaging, IEEE Transactions on 28*, 4 (2005), 694–703.

[82] Palankovski, V., and Selberherr, S. Thermal models for semiconductor device simulation. In *High Temperature Electronics, 1999. HITEN 99. The Third European Conference on* (1999), IEEE, pp. 25–28.

[83] Park, J.-H., Heriz, V. M., Shakouri, A., and Kang, S.-M. Ultra fast calculation of temperature profiles of VLSI ICs in thermal packages considering parameter variations. In *IMAPS, 40th International Symposium on Microelectronics* (2007).

[84] Park, J.-H., Shakouri, A., and Kang, S. Fast evaluation method for transient hot spots in VLSI ICs in packages. In *Quality Electronic Design, 2008. ISQED 2008. 9th International Symposium on* (2008), IEEE, pp. 600–603.

[85] Park, J.-H., Shakouri, A., and Kang, S.-M. Fast thermal analysis of vertically integrated circuits (3-D ICs) using power blurring method. In *ASME 2009 InterPACK Conference collocated with the ASME 2009 Summer Heat Transfer Conference and the ASME 2009 3rd International Conference on Energy Sustainability* (2009), American Society of Mechanical Engineers, pp. 701–707.

[86] Park, J.-H., Wang, X., Shakouri, A., and Kang, S.-M. Fast computation of temperature profiles of VLSI ICs with high spatial resolution. In *Semiconductor Thermal Measurement and Management Symposium, 2008. SEMI-THERM 2008. Twenty-Fourth Annual IEEE* (2008), IEEE, pp. 50–54.

[87] Poppe, A., Farkas, G., Parry, J., Szabó, P., Rencz, M., and Székely, V. DELPHI style compact modeling of stacked die packages. In *Semiconductor Thermal Measurement and Management Symposium, 2007. SEMI-THERM 2007. Twenty Third Annual IEEE* (2007), IEEE, pp. 248–254.

[88] Rencz, M. New possibilities in the thermal evaluation, offered by transient testing. *Microelectronics journal 34*, 3 (2003), 171–177.

[89] Rencz, M., and Székely, V. Non-linearity issues in the dynamic compact model generation [package thermal modeling]. In *Semiconductor Thermal Measurement and Management Symposium, 2003. SEMI-THERM 2003. Ninteenth Annual IEEE* (2003), IEEE, pp. 263–270.

[90] Rencz, M., and Székely, V. Studies on the nonlinearity effects in dynamic compact model generation of packages. *Components and Packaging Technologies, IEEE Transactions on 27*, 1 (2004), 124–130.

[91] Rencz, M., Székely, V., and Poppe, A. A methodology for the co-simulation of dynamic compact models of packages with the detailed models of boards. *Components and Packaging Technologies, IEEE Transactions on 30*, 3 (2007), 367–374.

[92] Roache, P. J. Quantification of uncertainty in computational fluid dynamics. *Annual Review of Fluid Mechanics 29*, 1 (1997), 123–160.

[93] Roache, P. J. *Verification and Validation in Computational Science and Engineering*. Hermosa Pub, 1998.

[94] Schütze, T. Thermal equivalent circuit models, Application note, Infineon, 2008. `http://www.infineon.com/dgdl/Infineon-AN2008_03_Thermal_equivalent_circuit_models-AN-v1.0-en.pdf?fileId=db3a30431a5c32f2011aa65358394dd2`.

[95] Schweitzer, D. Thermal transient multisource simulation using cubic spline interpolation of Zth functions. *arXiv preprint arXiv:0709.1852* (2007).

[96] Schweitzer, D., and Chen, L. Heat spreading revisited–effective heat spreading angle. In *Thermal Measurement, Modeling & Management Symposium (SEMI-THERM), 2015 31st* (2015), IEEE, pp. 88–94.

[97] Shah, M., Barren, J., Brooks, J., Golla, R., Grohoski, G., Gura, N., Hetherington, R., Jordan, P., Luttrell, M., Olson, C., et al. UltraSPARC T2: A highly-treaded, power-efficient, SPARC SOC. In *Solid-State Circuits Conference, 2007. ASSCC'07. IEEE Asian* (2007), IEEE, pp. 22–25.

[98] Sheahan, D. OpenSPARC T2, Overview (presentation), 2008. http://www.oracle.com/technetwork/systems/opensparc/2008-oct-opensparc-slide-cast-05-ds-1539006.html.

[99] Sodan, V., Stoffels, S., Oprins, H., Baelmans, M., Decoutere, S., and De Wolf, I. A modeling and experimental method for accurate thermal analysis of AlGaN/GaN powerbars. In *Power Semiconductor Devices & IC's (ISPSD), 2015 IEEE 27th International Symposium on* (2015), IEEE, pp. 377–380.

[100] Sridhar, A., Vincenzi, A., Ruggiero, M., Brunschwiler, T., and Atienza, D. 3D-ICE: Fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling. In *Computer Aided Design, 2010. ICCAD-2010. IEEE/ACM International Conference on* (2010), IEEE Press, pp. 463–470.

[101] Stan, M. R., Skadron, K., Barcella, M., Huang, W., Sankaranarayanan, K., and Velusamy, S. HotSpot: A dynamic compact thermal model at the processor-architecture level. *Microelectronics Journal 34*, 12 (2003), 1153–1165.

[102] Swinnen, B., Jourdain, A., De Moor, P., and Beyne, E. *Wafer Level 3-D ICs Process Technology*. Springer, 2008, ch. Direct Hybrid Bonding, in Wafer Level 3-D ICs Process Technology, pp. 257–267.

[103] Swinnen, B., Ruythooren, W., De Moor, P., Bogaerts, L., Carbonell, L., De Munck, K., Eyckens, B., Stoukatch, S., Tezcan, D. S., Tokei, Z., et al. 3D integration by Cu-Cu thermo-compression bonding of extremely thinned bulk-Si die containing 10 μm pitch through-Si vias. In *Electron Devices Meeting, 2006. IEDM'06. International* (2006), IEEE, pp. 1–4.

[104] Szekely, V. *Nonlinear and Distributed Circuits*. CTC, Taylor & Francis, 2005, ch. Distributed RC networks, in Nonlinear and Distributed Circuits.

[105] Torregiani, C., Oprins, H., Vandevelde, B., Beyne, E., and De Wolf, I. Compact thermal modeling of hot spots in advanced 3D-stacked ICs. In *Electronics Packaging Technology Conference, 2009. EPTC'09. 11th* (2009), IEEE, pp. 131–136.

[106] Van Loan, C. *Computational Frameworks for the FFT*. SIAM, 1992.

[107] von Trapp, F. 3D Integration: A progress report, Semiconductor equipment and materials international, 2009. http://www.semi.org/cms/groups/public/documents/web_content/ctr_033139.pdf.

[108] Wang, B., and Mazumder, P. Fast thermal analysis for VLSI circuits via semi-analytical Green's function in multi-layer materials. In *Circuits and Systems, 2004. ISCAS'04. Proceedings of the 2004 International Symposium on* (2004), vol. 2, IEEE, pp. II–409.

[109] Wang, B., and Mazumder, P. Accelerated chip-level thermal analysis using multilayer Green's function. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on 26*, 2 (2007), 325–344.

[110] Weast, R. C., Astle, M. J., and Beyer, W. H. *CRC handbook of chemistry and physics : a ready-reference book of chemical and physical data*. Boca Raton, Fla: CRC Press, 1984.

[111] Wgsimon. Transistor count and Moore's law - 2011. `http://commons.wikimedia.org/wiki/File:Transistor_Count_and_Moore's_Law_-_2011.svg#file`.

[112] Yovanovich, M., Muzychka, Y., and Culham, J. Spreading resistance of isoflux rectangles and strips on compound flux channels. *Journal of Thermophysics and Heat Transfer 13*, 4 (1999), 495–500.

[113] Zhan, Y., and Sapatnekar, S. S. High-efficiency Green function-based thermal simulation algorithms. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on 26*, 9 (2007), 1661–1675.

[114] Zhang, Y., Dembla, A., Joshi, Y., and Bakir, M. S. 3D stacked microfluidic cooling for high-performance 3D ICs. In *Electronic Components and Technology Conference (ECTC), 2012 IEEE 62nd* (2012), IEEE, pp. 1644–1650.

[115] Zhou, P., Ma, Y., Li, Z., Dick, R. P., Shang, L., Zhou, H., Hong, X., and Zhou, Q. 3D-STAF: scalable temperature and leakage aware floorplanning for three-dimensional integrated circuits. In *Computer-Aided Design, 2007. ICCAD 2007. IEEE/ACM International Conference on* (2007), IEEE, pp. 590–597.

[116] Ziabari, A. Fast static and dynamic grid level thermal simulation considering temperature dependent thermal conductivity of silicon. Master's thesis, University of California, Santa Cruz, 2012.

[117] Ziabari, A., Bian, Z., and Shakouri, A. Adaptive power blurring techniques to calculate IC temperature profile under large temperature variations. In *International Microelectronic Assembly and Packaging Society* (2010).

[118] Ziabari, A., and Shakouri, A. Fast thermal simulations of vertically integrated circuits (3D ICs) including thermal vias. In *Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2012. 13th IEEE Intersociety Conference on* (2012), pp. 588–596.

[119] Zienkiewicz, O., and Taylor, R. *The Finite Element Method, 4th Edition*. McGraw Hill, New York, 1987.

Curriculum vitae

Federica Lidia Teresa Maggioni
°16 March 1987, Monza (Italy)

Education

May 2012 - present

PhD student, Department of Mechanical Engineering, KULeuven & REMO, imec (Leuven, Belgium)
Supervisor: Prof. dr. ir. Martine Baelmans

September 2009 - July 2011

Master in Applied Mathematics, Università degli studi di Milano (Milano, Italy)

Master thesis: "Analysis and Numerical Simulation of Dielectrophoretic Experiments in Silicon Lab-On-Chip Devices"
Supervisor: Prof. dr. Giovanni Naldi

September 2009 - July 2011

ECMI diploma, European Programme in Mathematics for Industry, Technomathematics curriculum.

September 2009 - March 2010

Erasmus student, Chalmers University of Technology (Göteborg, Sweden)

September 2006 - July 2009

Bachelor in Applied Mathematics, Università degli studi di Milano (Milano, Italy)

Bachelor thesis: "Dinamica di Sistemi Planetari Extrasolari"
Supervisor: Prof. dr. Antonio Giorgilli

Professional experience

May 2012 - present

PhD researcher, Department of Mechanical Engineering, KULeuven & REMO, imec (Leuven, Belgium)

September 2011 - April 2012

Business Intelligence Consultant, Sopra Group (Assago, Italy)

October 2010 - July 2011

Intern, Master Thesis project, STMicroelectronics (Agrate Brianza, Italy)

List of publications

[1] **Maggioni, F.**, Oprins, H., Beyne, E., DeWolf, I., and Baelmans, M. Convolution based compact thermal model for 3D-ICs: methodology and accuracy analysis. In *Thermal Investigations of ICs and Systems, 2013. THERMINIC 2013. 19th International Workshop on* (2013), IEEE, pp. 152– 157.

[2] **Maggioni, F.**, Oprins, H., Beyne, E., De Wolf, I., and Baelmans, M. Fast convolution based thermal model for 3D-ICs: methodology, accuracy analysis and package impact. *Microelectronics Journal 45*, 12 (2014), 1746–1752.

[3] **Maggioni, F.**, Oprins, H., Beyne, E., De Wolf, I., and Baelmans, M. Convolution based steady state compact thermal model for 3D integrated circuits: Methodology for including the thermal impact of die to die interconnections. In *Proceedings of the 15th International Heat Transfer Conference, IHTC-15 August 10-15, 2014, Kyoto, Japan* (2014), Begell House.

[4] **Maggioni, F.**, Oprins, H., Beyne, E., De Wolf, I., and Baelmans, M. Convolution based compact thermal model application to the evaluation of the thermal impact of die to die interface including interconnections. In *Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2014. 14th IEEE Intersociety Conference on* (2014), IEEE, pp. 98–106.

[5] **Maggioni, F.**, Oprins, H., Milojevic, D., Beyne, E., De Wolf, I., and Baelmans, M. 3D-Convolution based fast transient thermal model for 3D integrated circuits: Methodology and applications. In *Thermal Measurement, Modeling & Management Symposium (SEMI-THERM), 2015 31st* (2015), IEEE, pp. 107–112.

[6] Oprins, H., Cherman, V., Van der Plas, G., **Maggioni, F.**, De Vos, J., Wang, T., Daily, R., and Beyne, E. Experimental thermal characterization and thermal model validation of 3D packages using a programmable thermal test chip. In *Electronic Components and Technology Conference (ECTC), 2015 IEEE 65th* (2015), IEEE, pp. 1134–1141.

[7] Oprins, H., Cherman, V., Van der Plas, G., **Maggioni, F.**, De Vos, J., and Beyne, E. Thermal experimental and modeling analysis of high power 3D

packages. In *IC Design & Technology (ICICDT), 2015 International Conference on* (2015), IEEE, pp. 1–4.

[8] **Maggioni, F.**, Oprins, H., Beyne, E., De Wolf, I., and Baelmans, M. Fast transient convolution based thermal modeling methodology for including the package thermal impact in 3D-ICs. *Components, Packaging, and Manufacturing Technology, IEEE Transactions on 6*, 3 (2016), 424-431.

[9] Oprins, H., **Maggioni, F.**, Cherman, V., Van der Plas, G. and Beyne, E. *Thermal Modeling and Experimental Model Validation for 3D Stacked ICs*, in Paul D. Franzon, Eric J. Marinissen and Muhannad S. Bakir (Eds.), *Handbook of 3D Integration – Volume 4: 3D Design, Test, and Thermal Management*, Wiley-VCH, 2016.

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF MECHANICAL ENGINEERING
APPLIED MECHANICS AND ENERGY CONVERSION SECTION
Celestijnenlaan 300 - box 2421
B-3001 Leuven (Belgium)
FedericaLidiaTeresa.Maggioni@kuleuven.be
http://www.mech.kuleuven.be/en/tme/